Erkenn
DOI 10.1007/s10670-016-9862-2

ORIGINAL RESEARCH

Vagueness and Aggregation in Multiple Sender Channels

Jonathan Lawry[1] · **Oliver James**[1]

Received: 23 July 2015 / Accepted: 9 November 2016

Abstract Vagueness is an extremely common feature of natural language, but does it actually play a positive, efficiency enhancing, role in communication? Adopting a probabilistic interpretation of vague terms, we propose that vagueness might act as a source of randomness when deciding what to assert. In this context we investigate the efficacy of multiple sender channels in which senders choose assertions stochastically according to vague definitions of the relevant words, and a receiver then aggregates the different signals. These vague channels are then compared with Boolean channels in which assertions are selected deterministically based on classical (crisp) definitions. We show that given a sufficient number of senders, a linear stochastic channel outperforms Boolean channels when performance is measured by the expected squared error between the actual value described by the senders and the receiver's estimate of it based on the signals they receive. The number of senders required for vague channels to be at least as accurate as Boolean channels is shown to be a decreasing function of the size of the language i.e. the number of description labels available to the senders. Vague channels are then shown to be robust to transmission error provided the error rate is not too large. In addition, we investigate the behaviour of both Boolean and vague channels for a parametrised family of distributions on the input values. Finally, we consider optimal vague channels assuming a fixed number of senders and show that, provided there are more than two senders, a vague channel can be found that outperforms the optimal Boolean channel. In this context, we show that for channels with relatively low numbers of senders S-curve production functions are optimal.

✉ Jonathan Lawry
j.lawry@bris.ac.uk

[1] Department of Engineering Mathematics, University of Bristol, Bristol BS8 1UB, UK

Published online: 22 March 2017

1 Introduction

Vagueness is ubiquitous in natural language, but it is unclear what practical role, if any, it plays in our communication. For example, is the vagueness of adjective definitions an efficiency enhancing feature of the way in which we represent concepts, or is it an unfortunate, if perhaps inevitable, side-effect of the way in which language is acquired, or has evolved (O'Connor 2013)? A fundamental difficulty encountered by any general attack on this problem relates to the breadth of the concept of vagueness itself. Vagueness is a multi-faceted phenomenon and although it is clearly different from ambiguity and imprecision there are still differing opinions as to exactly what linguistic phenomena come under its umbrella. Keefe and Smith (2002) identify three interrelated properties of vague predicates; (1) borderline cases (2) blurred boundaries and (3) susceptibility to sorites paradoxes. There is a subtle but important distinction between (1) and (2) which suggests that we may need to look at different aspects of communication in order to understand the possible utility of these different properties of vagueness. Explicit borderline cases are those which are neither members of a given category nor of its complement. Proposed models of this characteristic either permit truth gaps (Fine 1975), i.e. statements which are neither true nor false, or introduce a third truth-value to represent 'borderline' (Kleene 1952). van Deemter (2009a) has identified a number of communication scenarios in which vagueness can play a positive role including, for example, by mitigating the risk associated with making predictions or promises. Lawry and Tang (2012) suggested that borderline cases may indeed have a positive role to play in this form of risk management. The underlying intuition is that the presence of borderline cases provides additional flexibility within a payoff model when there is uncertainty about the possible outcomes. For instance, we might assume that the payoff from making a forecast that turns out to be borderline will lie somewhere between the payoffs from a forecast which turn out to be false and one which turns out to be true respectively. This extra flexibility allows agents to balance the vagueness of assertions against their uncertainty so as to maximise the expected payoff from making a forecast or a promise. Blurred boundaries on the other hand arise from a type of uncertainty about where exactly the boundary of a category lies, and we will argue below that this can be modelled probabilistically. In this paper we focus on the utility of blurred boundaries and, by adopting a probabilistic interpretation, we attempt to describe a communication scenario in which stochastic behaviour, resulting from vague definitions of adjectives along a continuous scale, is on average better than the optimal Boolean, i.e. non-vague, alternative.

Signalling games (Lewis 1969) have provided a common formalism in which to study the utility of vagueness, and in particular blurred boundaries, in communication [see van Deemter (2009b) for an overview of recent work]. Such games typically involve two agents, a sender and receiver, with a shared vocabulary consisting of a finite set of words, used to describe an underlying reality of which the sender but not the receiver has direct knowledge. Each agent then adopts transmission and interpretation strategies so as to maximize their respective utilities.

For example, De Jaegher (2003) investigates the role of vagueness in signalling games in which the sender and receiver have different and possibly conflicting utilities. From an alternative perspective Franke et al. (2011) suggest that vagueness is a natural property for boundedly rational agents. In particular, they consider the cases in which agents have bounded rationality due to memory limitations and also due to random error i.e. noise. Perhaps the most compelling study of vagueness in communication, however, is the still unpublished work of Lipman (2009) in which signalling is studied assuming the Gricean maxim (1975) that both sender and receiver aim to communicate as effectively as possible. Lipman's result shows that for rational agents, using vague definitions is always sub-optimal in comparison to a Boolean alternative. More specifically, vagueness is associated with the use of mixed strategies; these being probability distributions over pure strategies.[1] Informally stated, Lipman's main result is that no non-trivial mixed strategy is ever strictly better performing than any of the pure strategies to which it allocates non-zero probability. In other words, strict Nash equilibria will only contain pure strategies. One feature which is common to all of these studies is that in any signalling game there is only one sender. This immediately rules out the possibility of any form of information aggregation on the part of the receiver. We will now argue that it is exactly as part of such an aggregation process that labels with blurred boundaries may have some utility. We begin by considering a simple example.

There has been a street robbery in central Bristol. Around midday, a robber has approached a member of the public and stolen some money and their mobile phone. Due to the location and time at which the robbery took place, there are many witnesses, each able to provide a good description of the robber. The police officer in charge of the investigation takes formal statements in which the witnesses are asked to describe different characteristics of the robber including about their height. Now we have a clear intuition that the police officer benefits from having multiple statements, and to some extent, the more the better. This is no doubt partly because the different witnesses bring different perspectives, fill in the gaps left by others, and hence together provide a more complete overall picture of events. However, in addition, we suggest that an element of randomness on the part of witnesses in their choice of words can also provide the police officer with additional information. Furthermore, we suggest that the blurred boundaries or gradedness of vague words can be a natural source of this type of stochasticity. Suppose for simplicity that height is only describable using the two labels, *short* and *tall*, then if all witnesses describe the robber as *short*, then the police officer might infer that they are likely to be a prototypical short person. On the other hand, a 50–50 split between those witnesses who say *short* and those who say *tall* is more likely to suggest a person of intermediate height. Now notice that if instead of making stochastic assertions based on some form of graded concept definition, the witnesses were simply applying Boolean definitions of *short* and *tall*, then inference of this form would not be possible. To see this, suppose that all the witnesses share the same Boolean definitions of *short* and *tall*, according to which all heights less than a threshold θ

[1] Pure strategies correspond to deterministic functions mapping from inputs to words sent, and from words received to actions.

are classified as *short*, and all heights greater than θ as *tall*. In this case, if we assume a noise free model in which everyone receives the same information, then no matter what the robber's height, either all the witnesses would describe him as *short* or all as *tall*. Of course, in practice the witnesses are likely to differ, even in the case that they all adopt the same Boolean model. For example, there would be natural variation in their perceptions and in the conditions and locations where they each saw the robbery take place e.g. witnessing it from different angles and in different light. However, we suggest that in addition to this natural variation there can be a positive role to play for stochasticity directly induced by the blurred boundaries of vague categories.

2 The Uncertain Threshold Model of Vagueness

Probabilistic approaches to vagueness have a history dating back to Black (1937), and include work by Loginov (1966), Hisdal (1988), Edgington (1997) and more recently Lawry (2008) and Lassiter (2011). These models tend to be strongly interrelated, see Dubois and Prade (1997), and for graded adjectives a common formulation is in terms of an uncertain threshold value defined on a particular measurement scale (Cresswell 1976). Consider, for example, the adjective *short* defined on a height scale corresponding to the positive real numbers. As outlined in Sect. 1, a simple Boolean model is characterised by a threshold value θ, all heights below which are classified as being *short*. In the case of vague concepts it is then proposed that blurred category boundaries result from uncertainty about the exact value of θ. Lawry (2008) refers to this as *semantic uncertainty* and argues that it can be naturally quantified in terms of subjective probabilities. Both Lawry (2008) and Lassiter (2011) suggest that semantic uncertainty is a likely consequence of the empirical way in which language is acquired. In this paper we propose that it may also underlie stochastic assertion decisions which can play a positive role in communication scenarios where some form of aggregation is involved. For instance, suppose that for a witness in our robbery example her uncertainty about the threshold θ, defining the adjective short, is quantified by the probability density f.[2] The probability that this witness would classify a robber of height x metres as being short, then corresponds to the probability that the threshold value θ is at least x. This provides a natural definition for the membership degree of x in the category short as follows:

$$\mu_{short}(x) = P(\theta \geq x) = \int_x^{\infty} f(\theta)\, \mathrm{d}\theta = 1 - F(x)$$

where F is the cumulative distribution function of f. Applying a stochastic assertion model the witness would then describe the robber as being short with probability

[2] A number of recent studies have conducted experiments into how well different parameterised models for f fit with data relating to adjective use in natural language. See for example Lassiter and Goodman (2013) and Qing and Franke (2014).

$\mu_{short}(x)$ and as being tall with probability $\mu_{tall}(x) = 1 - \mu_{short}(x)$. In the following section we propose a simple stochastic communication channel involving vague labels defined in terms of uncertain thresholds. We show that in such a channel, by aggregating the varying signals from sufficiently many stochastic senders, a receiver can on average obtain a better estimate of the input being described, than by using an optimal Boolean model.

The uncertain threshold model has clear similarities to the epistemic theory of vagueness as expounded by Williamson (1992, 1994), although there are also subtle but important differences. Williamson proposes that there is a precise but unknown, and possibly unknowable, boundary between the extension of a vague concept and that of its negation. From this perspective vagueness can be captured within the framework of classical logic, with properties such as the law of excluded middle and the law of non-contradiction being preserved. The model we propose, though sharing with the epistemic theory the basic premise that vagueness can be understood in terms of precise but uncertain boundaries, makes a fundamentally different assumption regarding the nature of these boundaries and how the uncertainty about them arises. In particular, the epistemic theory would seem to assume the existence of some objectively correct boundary threshold between, for example, *short* and *not short*. This assumption lies at the heart of one of the main criticisms of epistemicism in the literature, that it does not provide a satisfactory account of the relationship between the semantics and the use of language (Keefe and Smith 2002; Smith 2008). That is, it seems clear that the meaning of vague concepts are in large part determined by their use over time by a diverse population of communicators. But the role of the individual within the epistemic theory appears to be that of learning the meaning of already fixed boundaries, a task at which, according to Williamson (1994), they can only hope to have at best partial success. In contrast, following Lawry (2008) and D'Odorico and Bennett (2013), we propose a model in which individuals adopt an *epistemic stance* by *assuming* the existence of precise boundary thresholds about which they are uncertain, and where they quantify this uncertainty using probability. However, the epistemic stance is understood to be a modelling assumption on the part of language users, and there is no implication that precise thresholds have an independent existence beyond the models. From this perspective there is a clear account of how language use determines semantics through an emergent process resulting from multiple interactions between individuals, each adopting the epistemic stance and updating their semantics by conditioning within a probabilistic representational model as outlined above. Indeed there is a growing literature on agent-based simulation studies in which simple probabilistic models of concepts are shown to converge across a population (Steels 1997; Steels and Belpaeme 2005; Eyre and Lawry 2014). Nonetheless, one might ask of such approach, why do individuals choose to adopt the epistemic stance, as opposed to an alternative representational model, given that, as admitted, there is no claim as to the objective existence of precise boundaries? A pragmatic response would be to claim that, faced with the challenge of deciding what to assert and of interpreting the assertions of others in a variety of contexts, individuals simply find it useful as part of a decision making and learning strategy to assume that there is a clear divide between those labels which are and those which

are not appropriate to assert. This is consistent with Lassiter's view that, rather than language being a simple precise entity, there are in fact a number of precise interpretations that can be employed in a given context (Lassiter 2011). Adopting a probabilistic approach in which individuals attempt to take account of their prior knowledge of language conventions and their models of other language users in order to choose between these various interpretations, could then be a natural way of bringing to bear already established tools for dealing with epistemic uncertainty when deciding between competing possible assertions. In this paper we furthermore propose that probabilistic definitions can also be exploited by communicating agents as a mechanism for generating stochastic uncertainty which then has a positive role to play in the aggregation of information from different signals.[3]

We should note that for some, even this pragmatic epistemic approach to vagueness may still be unpalatable. Hence, in the context of the current paper it is worth pointing out that the stochastic channels proposed in the sequel are also relevant for probabilistic but non-epistemic theories and even for non-probabilistic degree-based treatments of vagueness. To make the case for the former we consider the non-epistemic probabilistic approaches to vagueness as proposed by Borel, discussed and developed by Egré and Barberousse (2014), Egré (2016), and Kamp (1975). Borel applies statistical methods after identifying two main sources of variation in the way that individuals apply vague terms. For example, suppose that a witness' decision as to whether or not to describe the robber as short depends both on her perception of his height and on a precise threshold, with the term short being used provided that the former is less than the latter.[4] Variation in the responses of different witnesses then occurs as a result both of differences in their perceptions of the height of the robber and in the height thresholds they apply. Certainly the former would tend to naturally occur due to the inherent imperfection of human perception and other environmental effects e.g. differences in relative position, lighting etc. For

[3] The sorites paradox has traditionally been central to the study of vagueness. Several probabilistic accounts of sorites have been given in the literature, including Edgington (1997) and most recently Lassiter and Goodman (2015). These adopt two main interpretations of conditional rules; the material conditional interpretation (MC) and the probability conditional interpretation (PC) (Lassiter and Goodman 2015). For the probability threshold model outlined in this paper we can illustrate MC and PC as follows: Suppose we have a sequence of heights $\{x_k\}_{k=0}^{n}$ where $x_{k+1} = x_k + h$ for some small value $h > 0$. Further suppose that x_0 is clearly short while x_n is clearly not short. Now consider the sequence of sorites rules; IF $short(x_k)$ THEN $short(x_{k+1})$ for $k = 0, \ldots, n-1$. For MC each rule is held to be true with probability $1 - P(short(x_k) \wedge \neg short(x_{k+1})) = 1 - P(x_k \leq \theta < x_k + h) = 1 - F(x_k + h) - F(x_k)$. For PC the relevant probabilities are $P(short(x_{k+1})|short(x_k)) = \frac{P(\theta \geq x_{k+1})}{P(\theta \geq x_k)} = \frac{F(x_k+h)}{F(x_k)}$. Therefore, for both MC and PC the premises of sorites are compelling provide that $P(\theta \geq x_0)$ is high, $P(\theta \geq x_n)$ is low, and also that the probabilities of each of the conditionals is high for $k = 0, \ldots, n-1$. In fact it is straightforward to set up the sorites scenario so that this indeed the case. For instance, by taking F to be the cumulative distribution function of a normal distribution on θ with a suitable mean and variance, we can obtain that $\mu_{short}(x_0) \approx 1, \mu_{short}(x_n) \approx 0$ and that, for h sufficiently small, both MC and PC will result in conditional rules with probabilities close to 1 for $k = 0, \ldots, n-1$.

[4] One might be suspicious that by using a crisp threshold in this way Borel is implicitly endorsing the epistemic theory. However, Egré (2016) argues that this threshold should be thought of as forming part of 'a subjective decision rule' rather than as being an objective cut-off value. In fact, in essence this seems close to our interpretation of the epistemic stance as being a modelling assumption adopted by individuals to help them make decisions about assertions.

the latter Egré (2016) suggests that decision thresholds might be based on a representative value for the reference or context class, so for the robber example this could be the mean height of UK males, but with subjective differences between individuals about how exactly the threshold is derived from this value. Given this set-up we can try to understand the use of the term short in this particular context by running a controlled experiment in which a sample of individuals (witnesses) are shown a number of suspects with varying heights and asked whether or not they would describe them as being short, yes or no. From the resulting data statistical methods can then be employed so as to estimate a probability function quantifying the probability that the adjective short will be applied to describe someone of a given height. Now the fundamental difference between this statistical approach and the probabilistic model we have outlined above is that in Borel's approach probability describes the macro-level use of vague predicates across a population, capturing natural variations between individuals,[5] whilst we have proposed that each individual adopts a probabilistic model when deciding whether or not a vague term can be applied. However, stochastic channels as described below are agnostic as to the exact source of the variations between senders. Indeed for Borel's statistical model the main claim of our paper can be reformulated as follows; variation in the application of vague predicates with certain overall probabilistic profiles, can be a positive benefit in multi-sender channels.

A different non-epistemic probabilistic approach is proposed by Kamp (1975) as an extension of the supervaluation theory of vagueness (Fine 1975) in which a probability measure is introduced to weight the different admissible precisifications of a predicate. The membership of a element in the extension of the predicate is then taken to be the measure of the set of precisifications that contain it. Now clearly this model can also act as a source of stochasticity if, for example, when deciding whether or not to describe the robber as short, each witness picks a precisification at random according to the probability weighting and then checks if the robber's height is contained in the particular extension of short that they have chosen. Finally, a general degree-based view of vagueness defines the membership of the extension of a predicate as a function into [0, 1], but where there is no probabilistic interpretation of this membership function (Smith 2008). Even for this non-probabilistic model, stochastic channels can still be relevant provided that assertion decisions are made by employing a threshold on membership functions. For example, a witness will assert that 'the robber is short' provided that the robber's value in the witness' membership function for short exceeds some threshold θ. If θ is chosen stochastically then the type of signal aggregation proposed below can still be applied, but where the information conveyed over the channel relates to the robber's membership value in short rather than to a direct estimate of their height.

An outline of the remainder of the paper is as follows: Sect. 3 introduces the optimal Boolean binary channel as well as a simple vague channel involving the aggregation of stochastic signals. Section 4 then compares these two channels in terms of the expected squared error between the actual input and the receiver's

[5] In this respect Borel's approach is similar to that of Black (1937) and to the voting interpretation of fuzzy logic (Lawry 1998).

estimate of it, under the assumption that inputs are uniformly distributed on [0, 1]. In Sect. 5 we consider both Boolean and vague channels involving multiple labels. Section 6 investigates the robustness of vague channels to transmission error. In Sect. 7 we consider the situation in which the input distribution is unknown so that the channels cannot be optimised for a particular prior. In particular, we compare how both channels perform under a range of different input distributions. Section 8 considers optimal vague channels for different numbers of senders and, in particular, will show that S-curve membership functions perform well for channels with relatively low numbers of senders. Finally, in Sect. 9 we give some discussion and conclusions.

3 Boolean and Vague Channels

We now introduce a simple model of binary communication involving aggregation, as exemplified by the robber story from Sect. 1. An input value x is drawn at random from the normalised scale [0, 1] according to a uniform distribution. Each of a number of senders then select a label from the message set $\mathcal{M} = \{L_1, L_2\}$ which they judge to be an appropriate description of x, and transmit this to a single receiver. The receiver then aggregates these signals in order to determine an estimate y, of the value of x. We assume that all agents, senders and receiver, share the same definition of the labels in $\mathcal{M}$. Furthermore, we adopt Grice's assumption (1975) that all the senders aim to describe x in such a way as to enable the receiver to determine the best possible estimate. We now consider the two cases in which the labels in $\mathcal{M}$ are defined according to the standard Boolean model and according to an uncertain threshold based vague model.

3.1 The Optimal Boolean Channel

For binary Boolean channels we adopt a general fixed threshold model in which L_1 corresponds to the interval $[0, \theta)$ and L_2 to $[\theta, 1]$, for some threshold value θ in [0, 1]. That is, any value $x < \theta$ is *always* described as L_1 and any $x \geq \theta$ is *always* described as L_2. As discussed in Sect. 1, in such cases the receiver does not benefit from multiple signals since, given a shared Boolean model, all senders will assert identical descriptions of x.[6] Consequently we can simplify any such Boolean channel so as to consist of only one sender S and a receiver R. The sender transmits either a 0 (i.e. $S = 0$) to stand for L_1 or a 1 (i.e. $S = 1$) to stand for L_2. The receiver then estimates x to be y_0, a typical L_1 value, if they receive a 0 and to be y_1, a typical L_2 value, if they receive a 1. Assuming that x is uniformly distributed on [0, 1] we can measure the accuracy of this channel by evaluating the expected value of

[6] Note that this is not the case if the individual senders are subject to independent identically distributed sensor noise. For example, if each sender perceives $x + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma)$ then employing multiple senders can indeed improve estimation performance for Boolean channels. Ribeiro and Giannakis (2006) consider this case in detail including the selection of optimal threshold parameters under different conditions.

$(x-y)^2$, which we denote by $\mathbb{E}^{\mathbf{B}}((x-y)^2)$. Unsurprisingly this value is minimal when $\theta = \frac{1}{2}, y_0 = \frac{1}{4}$ and $y_1 = \frac{3}{4}$.

Theorem 1 *For a Boolean channel, if L_1 is defined as the interval $[0, \theta)$ and L_2 as the interval $[\theta, 1]$ and*

$$y = \begin{cases} y_0 : R = 0 \\ y_1 : R = 1 \end{cases},$$

and assuming that x is uniformly distributed on [0, 1], then $\mathbb{E}^{\mathbf{B}}((x-y)^2)$ is minimal when $\theta = \frac{1}{2}, y_0 = \frac{1}{4}$ and $y_1 = \frac{3}{4}$.

3.2 A Multiple Sender Vague Channel

We now propose a multiple sender vague channel in which signals from a number of stochastic senders are aggregated by a receiver so as to estimate the input variable. In contrast to the Boolean channel, in the vague channel all senders and receiver adopt a probabilistic interpretation of the labels in $\mathcal{M}$ as described in Sect. 2. More formally, there are $n+1$ agents corresponding to n senders $S_1, \ldots, S_n$ and a receiver R. Given the same input $x \in [0, 1]$ each sender independently selects a message from the set $\mathcal{M} = \{L_1, L_2\}$ and transmits either a 0 (i.e. $S_j = 0$) standing for L_1 or a 1 (i.e. $S_j = 1$) standing for L_2. All agents adopt the same shared probabilistic definition of $\mathcal{M}$ in which L_1 is $[0, \theta)$ and L_2 is $[\theta, 1]$ and where θ is an uncertain threshold which we assume to be uniformly distributed on [0, 1].[7] This results in the membership functions $\mu_{L_1}(x) = 1 - x$ and $\mu_{L_2}(x) = x$. We then assume that for each sender S_j the choice of signal, either 0 or 1, is stochastic with $P(S_j = 0|x) = \mu_{L_1}(x) = 1 - x$, and $P(S_j = 1|x) = \mu_{L_2}(x) = x$ (see Fig. 1). R receives an n-bit sequence of 1's and 0's from the different senders, where R_j denotes the signal received from sender S_j. R then aggregates these signals in order to obtain an estimate y, of the input x (see Fig. 2). We initially adopt the simple frequency estimator;

$$y = \frac{T}{n} \quad \text{where} \quad T = \sum_{j=1}^{n} R_j$$

4 A Comparison of Boolean and Vague Binary Channels

Assuming that x is uniformly distributed on [0, 1] we can use elementary statistics to evaluation the expected squared error for the vague channel described in Sect. 3.2 and denoted $\mathbb{E}^{\mathbf{V}}((x-y)^2)$, as follows:

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) = \int_0^1 \mathbb{E}^{\mathbf{V}}((x-y)^2|x)\,\mathrm{d}x$$

Given input x, T is distributed according to a binomial distribution with parameters n and x. Hence, $\mathbb{E}(T|x) = nx$ and $\mathbb{E}^{\mathbf{V}}(y|x) = x$. Therefore,

[7] From the perspective of the epistemic theory this corresponds to the situation in which the speaker is completely uncertain about θ.

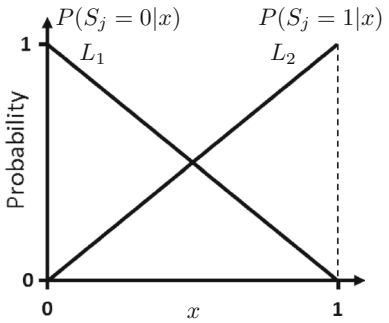

Fig. 1 Probabilities for sending a 0 or a 1 given x, derived from a vague definition of labels L_1 and L_2

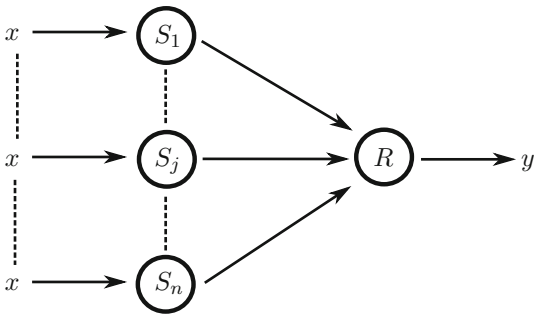

Fig. 2 A multiple sender vague channel

$$\mathbb{E}^{\mathbf{V}}((\mathbb{E}^{\mathbf{V}}(y|x) - y)^2|x) = \mathbb{V}^{\mathbf{V}}(y|x) = \mathbb{V}\left(\frac{T}{n}|x\right) = \frac{1}{n^2}\mathbb{V}(T|x) = \frac{x(1-x)}{n}$$

From this we obtain that:

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) = \int_0^1 \frac{x(1-x)}{n}\mathrm{d}x = \frac{1}{6n}$$

For the optimal Boolean channel we have instead that the expected squared error is given by:

$$\mathbb{E}^{\mathbf{B}}((x-y)^2) = \int_0^{\frac{1}{2}} \left(x - \frac{1}{4}\right)^2 \mathrm{d}x + \int_{\frac{1}{2}}^1 \left(x - \frac{3}{4}\right)^2 \mathrm{d}x = \frac{1}{48}$$

Now trivially, $\mathbb{E}^{\mathbf{V}}((x-y)^2)$ is a strictly decreasing function of n (see Fig. 3) and hence $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ provided that $n \geq 8$.

At this point we might be tempted to argue that a lower bound of 8 on the required number of senders does not make a strong case for the utility of vagueness in communication. After all how often do we have the luxury of aggregating assertions from that many different independent sources? However, note that we have not yet attempted to optimise the vague channel as we have done for the Boolean channel. We return to this issue in Sect. 8 where we show that there are vague channels that outperform the optimal Boolean channel when there are 2 or more senders. Initially, however, we investigate the behaviour of the linear vague channel described above (Fig. 1) as the number of labels in $\mathcal{M}$ increases and in this case show that the number of senders required to outperform the Boolean channel also decreases significantly. Furthermore, we then consider the robustness of the vague channel to noise and to ignorance about the underlying distribution on x.

5 Multiple Labels Channels

In this section we consider channels in which there are multiple labels so that $\mathcal{M} = \{L_1, \ldots, L_k\}$ for $k \geq 2$. We are thinking of these labels as representing higher granularity descriptions of values on some common underlying scale. For example, instead of simply describing the robber as being either *short* or *tall*, witnesses might instead choose between the three labels; *short*, *medium* and *tall*, or perhaps between the five labels; *very short*, *short*, *medium*, *tall* and *very tall*. As the number of labels increases then each label refers to a more and more specific range on the scale.

We assume that Boolean labels are defined in terms of $k+1$ fixed threshold values $0 = \theta_0 \leq \theta_1 \leq \ldots \leq \theta_{k-1} \leq \theta_k = 1$ such that the label L_i corresponds to the interval $[\theta_{i-1}, \theta_i)$ for $i = 1, \ldots, k-1$ and L_k corresponds to $[\theta_{k-1}, \theta_k]$. As in Sect. 4, the Boolean nature of this channel and the fact that the same label definitions are shared by all agents mean that we need only assume one sender and a receiver. The sender transmits a value in $\{0, \ldots, k-1\}$, where $S = i - 1$ stands for L_i, and upon receiving which the receiver estimates the value of x to be a typical value of L_i denoted by y_{i-1}. This form of channel fits within the general framework of quantization in multi-sensor platforms proposed by Gubner (1993). Gubner's model is more general in that, for example, it allows for different sensor reading from the different senders resulting from sensory noise and other environmental variations. From the following theorem we see that the expected squared error for this channel is minimal when the threshold values are regularly spaced between 0 and 1 and where the typical values are the mid points of each interval.

Theorem 2 *For a Boolean channel, if L_i is defined as the interval $[\theta_{i-1}, \theta_i)$ for $i = 1, \ldots, k-1$ and L_k is defined as $[\theta_{k-1}, \theta_k]$ where $0 = \theta_0 < \theta_1 < \ldots < \theta_{k-1} < \theta_k = 1$, and*

$$y = \{y_i : R = i \quad \text{for } i = 0, \ldots, k-1$$

then, assuming that x is uniformly distributed on $[0,1]$, $\mathbb{E}^{\mathbf{B}}((x-y)^2)$ is minimal when $\theta_i = \frac{i}{k}$ and $y_i = \frac{\theta_{i-1}+\theta_i}{2}$ for $i = 1, \ldots, k$.

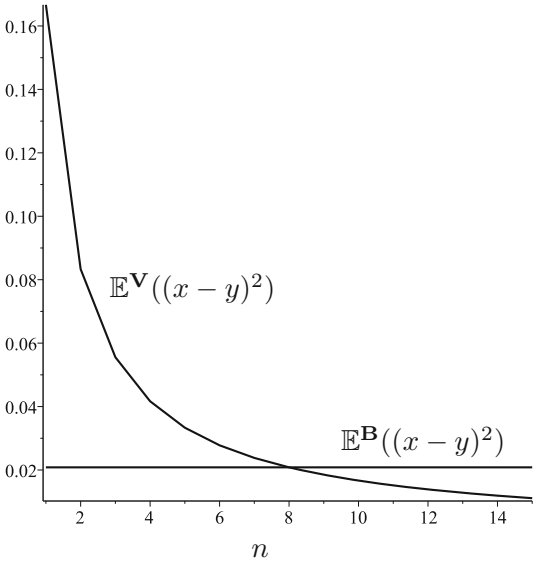

Fig. 3 $\mathbb{E}^{\mathbf{V}}((x-y)^2)$ and $\mathbb{E}^{\mathbf{B}}((x-y)^2)$ as functions of the number of senders n

For the vague channels with multiple labels we assume that the label L_i corresponds to the interval $[\theta_{i-1}, \theta_i)$ for $i = 1, \ldots, k-1$ but where each of the thresholds is uncertain.[8] There are many possible joint distributions on these $k-1$ thresholds satisfying the constraints that $\theta_{i-1} < \theta_i$, but here we adopt a simple formulation in which $\theta_i = \theta + \frac{i-1}{k-1}$ where the parameter θ is uniformly distributed on the interval $(0, \frac{1}{k-1})$. The memberships for the labels are then as follows (see Fig. 4):

$$\mu_{L_i}(x) = \begin{cases} (k-1)x - (i-1) : x \in \left[\dfrac{i-2}{k-1}, \dfrac{i-1}{k-1}\right) \\ \quad i - (k-1)x : x \in \left[\dfrac{i-1}{k-1}, \dfrac{i}{k-1}\right) \\ 0 : \text{otherwise} \end{cases} \qquad \text{for } i = 1, \ldots, k$$

Each of the n senders then stochastically transmits a value from $\{0, \ldots, k-1\}$ where $P(S_j = i-1|x) = \mu_{L_i}(x)$. R then receives a n-length sequence of numbers from $\{0, \ldots, k-1\}$ which they aggregate using the frequency estimator;

$$y = \frac{T}{n(k-1)} \quad \text{where} \quad T = \sum_{j=1}^{n} R_j$$

This form of multiple label linear vague channel is a special case of the model of probabilistic quantization proposed by Xiao et al. (2006). In Xiao et al. (2006) an upper bound on estimation error is determined for a sensor fusion platform employing linear probabilistic quantization and assuming that each sender is prone to independent noise drawn from a distribution with mean zero and a known

[8] We take $\theta_0 = 0, \theta_k = 1$ and L_k to be $[\theta_{k-1}, \theta_k]$.

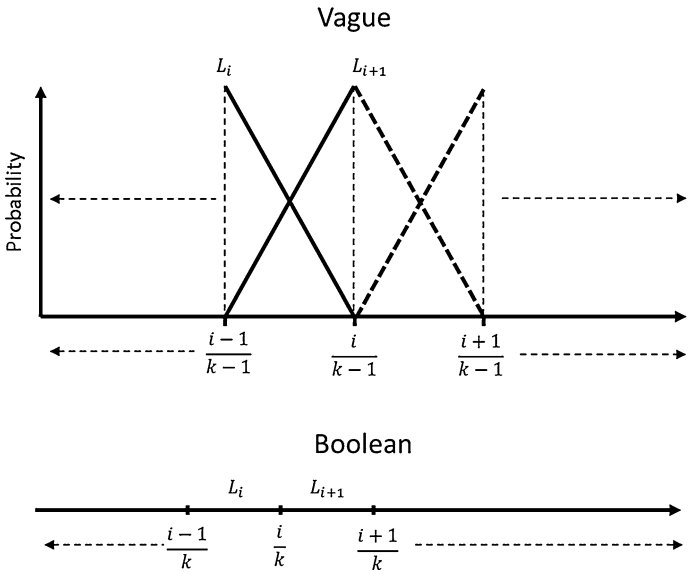

Fig. 4 Definition of vague and Boolean labels for a k label channel

standard deviation. Here, however, we focus on a direct comparison between stochastic channels of this kind and the optimal Boolean channel. The following results show that the minimal number of senders required for the vague channel with multiple labels to be on average at least as accurate as the comparable Boolean channel, is a decreasing function of the number of labels k (see Fig. 5). This value is strictly greater than 2 for all k, tending to 2 in the limit as k tends to infinity. In fact, for channels with 6 or more labels only 3 senders are required for the vague channel to be at least as accurate as the Boolean channel.

Lemma 3 *Let* $n_i = |\{j : S_j = i\}|$ *for* $i = 0, \ldots, k-1$. *If* $x \in [\frac{i-1}{k-1}, \frac{i}{k-1})$ *then*

$$y = \frac{\frac{n_i}{n} + i - 1}{k-1} = \frac{\frac{n_i}{n_{i-1}+n_i} + i - 1}{k-1}$$

Furthermore, $\mathbb{E}^{\mathbf{V}}(y|x) = x$.

Theorem 4 *If* x *is uniformly distributed on [0, 1] then* $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ *if and only if* $n \geq \left\lceil \frac{2k^2}{(k-1)^2} \right\rceil$.[9]

6 Robustness to Errors

It is commonly argued that systems which employ categories with fuzzy or blurred boundaries are inherently tolerant of errors due to the gradedness of category

[9] $\lceil z \rceil$ denotes the smallest natural number great than or equal to z.

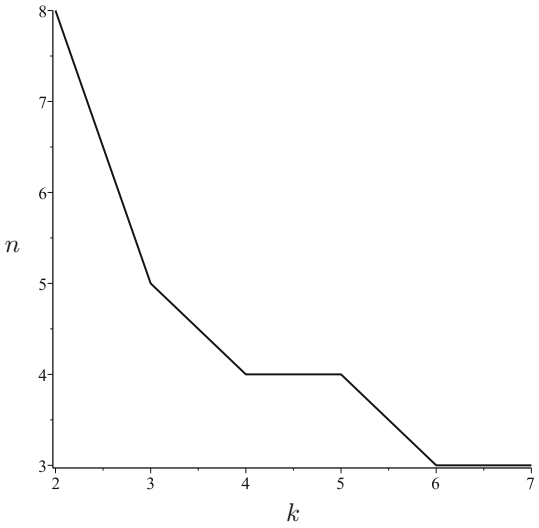

Fig. 5 The minimum number of senders n required such that $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ plotted as a function of the number of labels k

membership.[10] In our context we now investigate how tolerant binary vague channels are to transmission errors i.e. when $S_j \neq R_j$. For example, such errors could be due to the receiver mishearing the speaker in a noisy environment, or in our robbery example, information from a witness being misreported or misrecorded. Throughout this analysis we will compare the expected squared error of the vague channel to that of the error free optimal Boolean channel. For vague channels we consider the simple case in which there is a fixed probability α of an error occurring for each of the j channels to be aggregated. In other words;

$$P(R_j = 1|S_j = 0) = P(R_j = 0|S_j = 1) = \alpha \quad \text{for } j = 1, \ldots, n$$

The following result shows that provided the transmission error probability α is less than $\frac{1}{4}$ then by increasing the number of senders, vague channels can compensate for errors so as to still perform as well as the error free Boolean channel (see Fig. 6). As α tends to $\frac{1}{4}$ from below this minimum number of required senders tends to infinity. However, for example, to compensate for a 10% error rate only requires a relatively modest increase from 8 to 12 senders. Indeed, for small error probabilities of upto 0.045 only one additional sender is needed.

Theorem 5 *If x is uniformly distributed on [0, 1], and $\alpha < \frac{1}{4}$ then $\mathbb{E}^{\mathbf{V}}((x-y)^2|\alpha) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ if and only if $n \geq \left\lceil \frac{8(2\alpha - 2\alpha^2 + 1)}{1 - 16\alpha^2} \right\rceil$. If $\alpha \geq \frac{1}{4}$ then $\mathbb{E}^{\mathbf{V}}((x-y)^2|\alpha) \geq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ for all $n \geq 1$.*

[10] For example, see Hüllermeier (2011) for a discussion of the robustness of fuzzy methods used in machine learning.

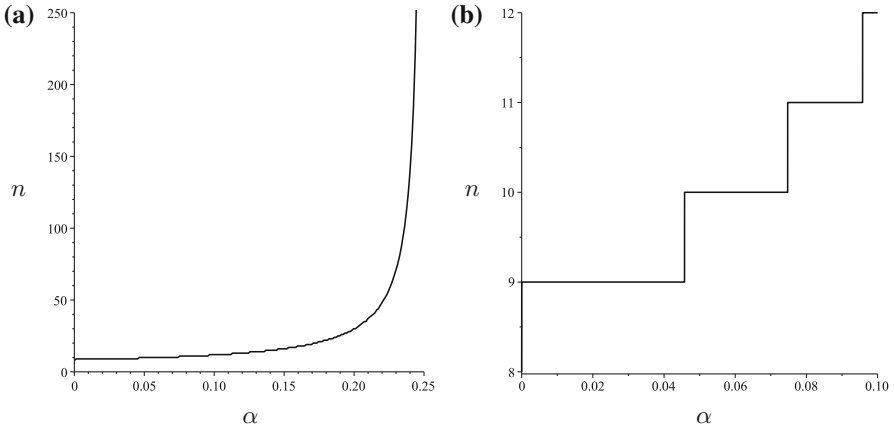

Fig. 6 The minimum value of n such that $\mathbb{E}^{\mathbf{V}}((x-y)^2|\alpha) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ plotted as a function of the channel error probability α. **a** α ranging from 0 to $\frac{1}{4}$. **b** α ranging from 0 to 0.1

7 Robustness to Ignorance

In the previous sections we have assumed that the distribution of the inputs x is known to be uniform on [0, 1]. Instead, we now consider the situation in which the distribution on inputs is unknown prior to communication so that it is not possible to a priori optimise the design of the channels in order to minimize expected squared error.[11] In the face of such ignorance we assess how the Boolean and vague channels introduced in Sect. 3 perform in different possible realities i.e. given different distributions on x. In the first instance we suppose that the world turns out to be such that inputs are symmetrically distributed about $\frac{1}{2}$. To model this scenario we evaluate the expected squared error for both channels assuming that x is distributed according to a symmetric beta distribution with parameter s i.e. with density function $\frac{x^{s-1}(1-x)^{s-1}}{\beta(s,s)}$ (see Fig. 7). The following result gives an expression for the minimal number of senders required for the vague channel to be at least as accurate as the Boolean channel as a function of the symmetric beta distribution parameter s. In the limit as s tends to infinity the required number of senders tend to 4. Furthermore, from Fig. 8 we can see that across all s the maximal number of required senders is 11. In other words, providing that the vague channel has at least 11 senders then we can be sure that it will be as least as accurate as the Boolean channel no matter what value of s characterises the true input distribution.

[11] Given ignorance of the distribution on inputs, agents might invoke the principle of insufficient reason and assume a uniform distribution. This would then motivate them to adopt a Boolean channel with $\theta = \frac{1}{2}, y_0 = \frac{1}{4}$ and $y_1 = \frac{3}{4}$ as described in Sect. 3.1.

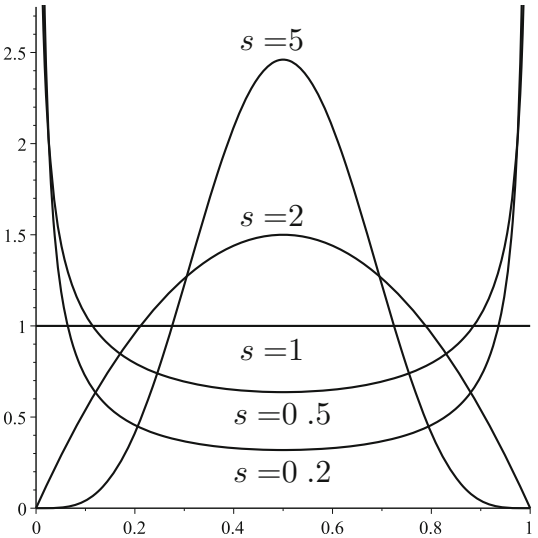

Fig. 7 Density functions for symmetric beta distributions with $s = 0.2, s = 0.5, s = 1, s = 2$ and $s = 5$

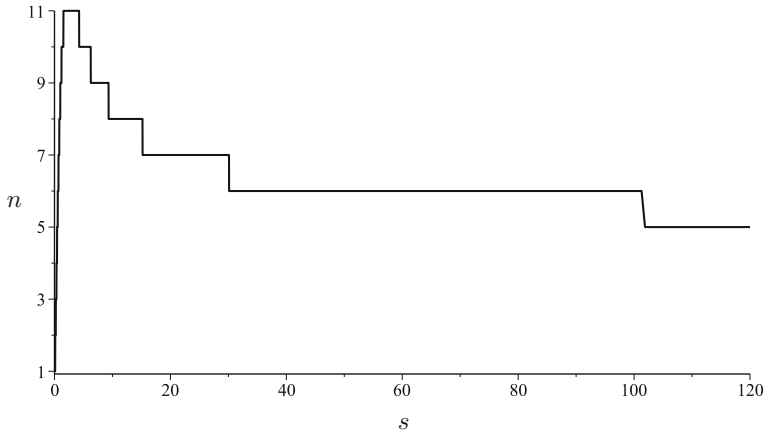

Fig. 8 The minimum number of senders n required such that $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$, assuming that x is distributed according to a symmetric beta distribution with parameter s, plotted as a function of s

Theorem 6 *If x is distributed according to a symmetric beta distribution with parameter $s > 0$, then $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ if and only if*

$$n \geq \left\lceil \frac{8s^2\beta(s,s)}{s(2s+5)\beta(s,s) - \left(\frac{1}{2}\right)^{2s+1} 16(2s+1)} \right\rceil$$

The assumption that inputs will turn out to be symmetrically distributed is of course a strong one, and may well be unrealistic. In order to investigate asymmetric input distributions we now evaluate the expected squared error for both channels

assuming that inputs follow a general beta distribution with parameters s and t i.e. with density function $\frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}$. The following result gives an expression for the minimum number of senders required for the vague channel under this distribution, as a function of the beta parameters s and t. From this we can obviously infer that no matter what values of s and t characterise the actual distribution of inputs there is always a minimum number of senders for which the vague channel is at least as accurate as the Boolean channel. Unfortunately, this minimal number of senders is unbounded as s and t vary. To see this consider the case where $s = 2t$. Figure 9b shows the beta density functions in this case for different values of t, all of which have an expected value of $\frac{3}{4}$. Furthermore, as t increases these density functions become increasingly peaked at $\frac{3}{4}$. Now clearly the Boolean channel will tend to be well suited to any such reality since the sender would be highly likely to transmit a 1, given which the receiver will estimate the value $y_1 = \frac{3}{4}$. Indeed Fig. 9a suggests that the minimum number of senders required for the vague channel given this family of skewed distributions is an unbounded strictly increasing function of t.

Theorem 7 *If x is distributed according to a beta distribution with parameters $s > 0$ and $t > 0$, then $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ if and only if*

$$n \geq \left\lceil \frac{16\beta(s,t)st}{8\beta\left(\frac{1}{2};s,t\right)(s-t)(s+t+1) - 16\left(\frac{1}{2}\right)^{s+t}(s+t+1) + (9t^2+9t-6st+s^2+s)\beta(s,t)} \right\rceil$$

8 Optimal Vague Channels

Up to this point we have focused on comparing a simple linear vague channel with the optimal Boolean channel for languages of different sizes, as well as under channel noise and when both senders and receives are ignorant about the underlying distribution of the input values. In this section we investigate the optimal vague channel for a fixed number of senders. To make a precise comparison between the optimal vague and Boolean channel we initially need to clarify what exactly we mean by *vague channel* in this more general context. From the discussion of the threshold model of vagueness in Sect. 2, we consider the labels $L_1 = [0, \theta)$ and $L_2 = [\theta, 1]$ where θ is a random variable with probability density function f and associated cumulative distribution F. We then have that S_j sends a 0 or 1 according to the generator function F as follows:

$$P(S_j = 0|x) = P(x < \theta) = 1 - F(x) \quad \text{and} \quad P(S_j = 1|x) = P(x \geq \theta) = F(x)$$

Now if we allow for the possibility that $f(x) = \delta(x - \frac{1}{2})$, i.e. the Dirac delta function at $\frac{1}{2}$, then this class of channels will also include the optimal Boolean channel. Hence, to make a clear distinction between vague and Boolean channels we insist that for vague channels f is a continuous function on [0, 1]. Given this requirement it follows that for channels with only one sender all vague channels have a strictly higher expected error than the optimal Boolean channel.

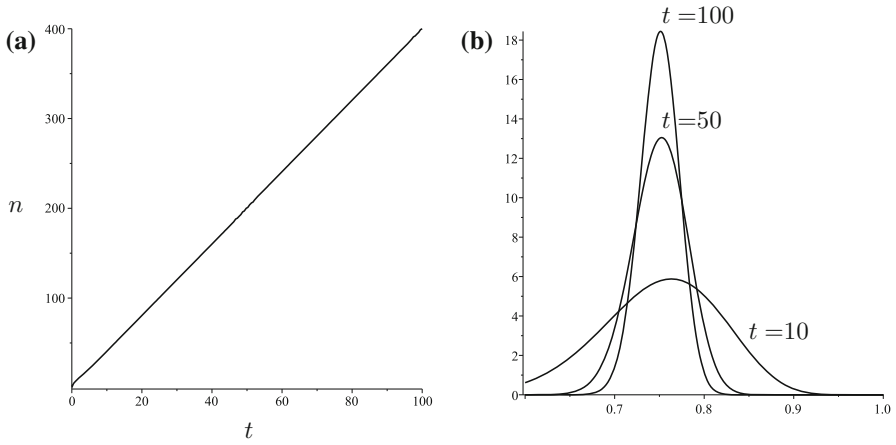

Fig. 9 The case in which x is distributed according an asymmetric beta distributions with parameters $2t$ and t. **a** The minimum value for n for which $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{B}^{\mathbf{B}}((x-y)^2)$. **b** Beta distribution with parameters $2t$ and t for $t = 10, t = 50$ and $t = 100$

Theorem 8 *There is no vague channel with only one sender such that* $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$.

In contrast, for $n \geq 2$ it is always possible to find a vague channel of this more general form which outperforms the optimal Boolean channel. However, the optimal distribution on the threshold θ will be different for different numbers of senders. To see this consider a vague channel with n senders and threshold cumulative distribution F then the error minimizing estimator of x from T is given by:

$$y = \mathbb{E}(x|T) = \int_0^1 xP(x|T)\,\mathrm{d}x = \frac{\int_0^1 xP(T|x)\,\mathrm{d}x}{\int_0^1 P(T|x)\,\mathrm{d}x} = \frac{\int_0^1 xF(x)^T(1-F(x))^{n-T}\,\mathrm{d}x}{\int_0^1 F(x)^T(1-F(x))^{n-T}\,\mathrm{d}x}$$

For example, if θ is uniformly distributed as in Sect. 3 then the error minimizing estimator of x corresponds to Laplace's rule so that $y = \frac{T+1}{n+2}$. In this case we obtain that $\mathbb{E}^{\mathbf{V}}((x-y)^2) = \frac{1}{6(n+2)}$ and hence, by using this estimator in place of the frequency $y = \frac{T}{n}$ the minimum number of senders for which $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ decreases from 8 to 6. More generally, we can also consider optimising the choice of threshold distribution F so as to minimise the expected error of the vague channel when applying the error minimizing estimator of x. Here we consider a parametrised family of density functions f in the form of normal distributions with mean $\frac{1}{2}$ and standard deviation σ, normalised so that all values of θ are between 0 and 1. In this case the cumulative distribution F has the following form:

$$F(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-\frac{1}{2}}{\sigma\sqrt{2}}\right) + \left(x-\frac{1}{2}\right)\left(1 + \mathrm{erf}\left(\frac{-\frac{1}{2}}{\sigma\sqrt{2}}\right)\right)\right.$$

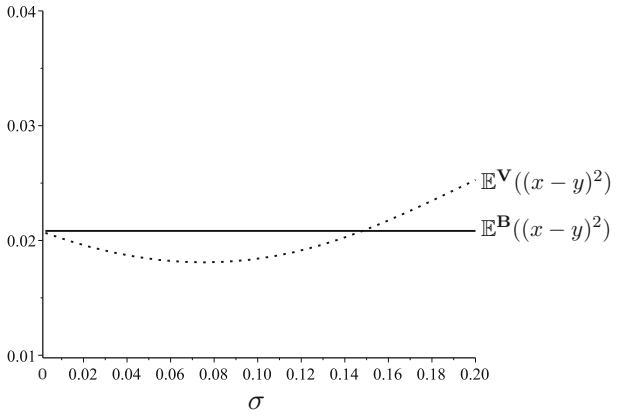

Fig. 10 $\mathbb{E}^{\mathbf{V}}((x-y)^2)$ and $\mathbb{E}^{\mathbf{B}}((x-y)^2)$ plotted against σ for a channel with 2 senders

Here we can view σ as a vagueness parameter such that as $\sigma \to 0$ then $F(x)$ tends to the step function so that the vague channel converges to the Boolean channel, whereas as $\sigma \to \infty$ then $F(x)$ tends to x giving the linear vague channel already investigated in this paper. Figure 10 shows $\mathbb{E}^{\mathbf{V}}((x-y)^2)$ for the error minimizing vague channel with two senders compared to $\mathbb{E}^{\mathbf{B}}((x-y)^2)$ as σ varies. The optimal two sender vague channel for this parametrised family of distributions is at $\sigma \approx 0.07532$ but the error minimising vague channel outperforms the optimal Boolean channel for $\sigma \leq 0.1482$. Note that the optimal distribution function is different for channels with different numbers of senders n. For example, Fig. 11 shows the optimal cumulative distributions and Fig. 12 shows the corresponding optimal values of σ for the channels with $n = 1, \ldots, 10$ senders. This suggests that vaguer channels are optimal when there are a larger numbers of senders, but that the gradient of this increasing trend in vagueness decreases with n. In terms of a direct comparison between vague and Boolean channels then for $n \geq 6$, adopting the error minimizing estimator of x ensures that $\mathbb{E}^{\mathbf{V}}((x-y)^2) < \mathbb{E}^{\mathbf{B}}((x-y)^2)$ for all $\sigma > 0$. For example, Fig. 13a, b shows the expected error for channels with 6 and 8 senders respectively plotted against σ and compared to the Boolean channel error.

9 Discussion and Conclusions

In this paper we have attempted to make the case for vague categories with blurred boundaries playing a positive role in a certain type of communication scenario in which a receiver aggregates signals from multiple senders. We have compared a simple vague channel with linear membership functions and frequency based aggregation to the optimal Boolean channel. Unsurprisingly, for error free channels in which the input distribution is a priori known to be uniform, the expected squared error for the vague channel is a strictly decreasing function of the number of senders. Since Boolean channels do not gain from having multiple senders then we

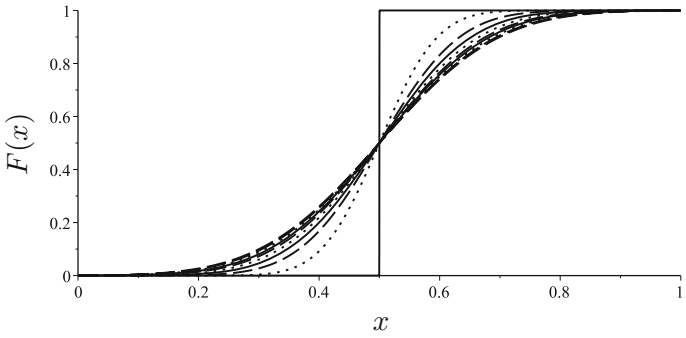

Fig. 11 Cumulative distribution F for the optimal channel for $n = 1, \ldots, 10$ senders

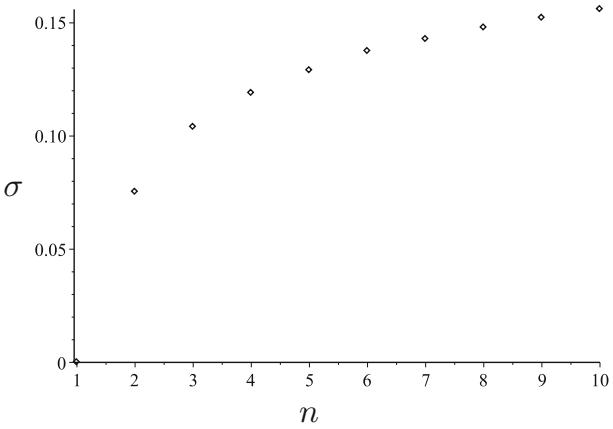

Fig. 12 Optimal values of σ for channels with $n = 1, \ldots, 10$ senders

can always identify a minimum number of senders above which the vague channel will be on average more accurate, in terms of expected squared error, than the comparable Boolean channel. Our focus has then been on identifying the minimal number of senders in different scenarios where there are multiple labels, channel error or prior ignorance about the input distribution, and also when optimal vague channels are considered. This is motivated by the intuition that the lower bound on the number of senders required by vague channels directly influences the strength of our case for the efficacy of blurred boundaries.

The plausibility of our argument that the blurred boundaries of vague predicates have a useful role to play as a natural source of stochastic assertion decisions, depends to a large part on the extent to which aggregation, of the form exemplified by our robbery story, is a common part of natural language communication. We note that for one sender and one receiver channels our results are entirely consistent with those of Lipman (2009), all be it formulated differently. Stochastic channels of the form we have proposed are undoubtedly suboptimal in such cases (see Theorem 8). For our argument in favour of vagueness to be in any way convincing it would need

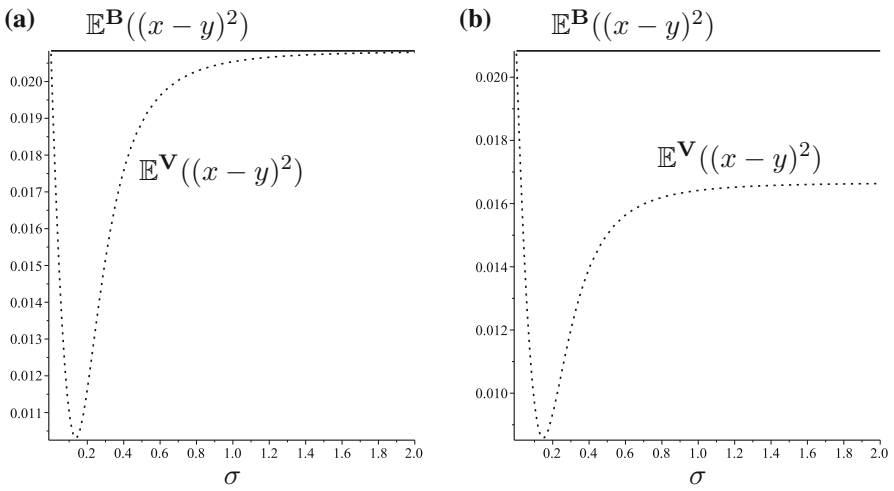

Fig. 13 $\mathbb{E}^{\mathbf{V}}((x-y)^2)$ and $\mathbb{E}^{\mathbf{B}}((x-y)^2)$ plotted against σ for channels with 6 and 8 sender channel. **a** Expected errors for a channel with 6 senders. **b** Expected errors for a channel with 8 senders

to hold that some level aggregation is a common part of linguistic communication, indeed even more common than one-on-one interactions of the type modelled by signalling games. We do not attempt to directly make this case here, neither are we aware of any empirical studies which look specifically into this claim. Instead, as we emphasised earlier, our goal is only to identify a possible scenario in which vagueness can be useful. However, it nonetheless seems clear that the larger the number of senders required for the vague channel to at least match the accuracy of the Boolean channel, the less compelling is the case for stochastic aggregation being a common feature of language. From this respect both our result for multiple vague channels (Theorem 4) and our study of optimal vague channels (Sect. 8) are both encouraging. For the former we have shown that the number of senders required for the linear vague channel to outperform the Boolean channel decreases rapidly as the number of labels increases (see Fig. 5). For the latter we have shown that by adopting the error minimizing estimator of x and then by selecting the distribution on θ from a parametrised family with mean $\frac{1}{2}$, we can identify a unique channel which minimizes the value of $\mathbb{E}^{\mathbf{V}}((x-y)^2)$ for any fixed number of senders. Furthermore, provided that $n \geq 2$ then vague channels can be found with a lower expected error than the optimal Boolean channel. Note that different vague channels are optimal for aggregating different numbers of senders, with more vague label definitions being preferred for larger n (see Fig. 12). Tantalisingly the type of S-curve reported in recent experimental studies on scalar adjectives (Lassiter and Goodman 2013; Qing and Franke 2014) are similar in form to those cumulative distribution functions optimal for channels with relatively low numbers of senders (see Fig. 11). This would then be consistent with the form of limited aggregation that one might expect to find in natural language where senders are scarce resources and where normally there will only be a small number of them. Certainly, the

accuracy gained by using vague channels could potentially confer a significant advantage to both senders and receivers. For instance, in multi-label communication with 7 labels and 3 senders the expected squared error for the vague channel is around 11% lower than that of the Boolean channel with the same number of senders. Certainly message sets with around 7 labels are not unrealistic, being consistent with the famous magic number theory of Miller (1956) which proposes bounds on the number of graduations on a numerical scale based on the limitations of human memory. Furthermore, a vague channel optimised for 2 labels and for only 2 senders has an error around 13% lower than the Boolean channel.

From a game theory perspective and with reference to Lipman (2009) we might wonder how the multi-sender games described in this paper can demonstrate the utility of stochasticity in communication and hence escape the general result that mixed strategies are always suboptimal to pure strategies. In order to reconcile Lipman's observation with our results we must first clarify the type of games being played by the Boolean and vague channels respectively. For instance, for the Boolean channel only two strategies are available to senders; transmit 0 or transmit 1. In contrast, for the vague channel we should think of the n senders as a compound aggregated sender $\mathbf{S}$ who can choose between signals $0, 1, \ldots, n$, i.e. the possible values of T, and whose available strategies is the set of all the binomial distributions on $\{0, \ldots, n\}$. Hence, for $n > 1$ one explanation for the superior performance of vague channels is that the sender simply has more strategies to chose from than in the Boolean channel. The question remains, however, why is a pure strategy not also optimal for the vague channel? The reason for this lies in the restricted set of strategies available to $\mathbf{S}$. In a mixed-strategy game $\mathbf{S}$ would be allowed to chose any strategy from Δ, the set of all probability distributions on $\{0, \ldots, n\}$ (Osborne and Rubinstein 1994). However, the set of binomial distributions is a non-convex strict subset of Δ. In particular, it does not include any pure strategy of the form $\mathbf{S} = T$, where $T \in \{1, \ldots, n-1\}$. However, in the case that $x \in (0, 1)$ it is exactly such a pure strategy, i.e. where $\mathbf{S} = \lceil nx \rceil$, that is optimal in the full mix-strategy game. On the other hand, permitting this optimal strategy would be hard to justify in the context of natural language communication, since it would require that the n senders collaborate so as to transmit the best n-bit approximation to x i.e. any combination of signals in which the number of ones is exactly $\lceil nx \rceil$. Essentially this would then be equivalent to a single n-bit channel, rather than n 1-bit channels. However, in natural language scenarios such as the robbery example in which the descriptions of a number of *independent* witnesses are aggregated, it is the latter which would seem to provide the more appropriate model.[12]

To assume error free channels for which the input distribution is completely known prior to communication, is unrealistic. However, we have shown that vague channels are robust to reasonable levels of transmission error i.e. with error probability less than $\frac{1}{4}$. In such cases by increasing the number of senders a vague

[12] In sensor networks it can be appropriate and useful to consider different models of collaboration between senders. For example, Luo (2005) proposes a scheme in which, while each sender transmits independently, they are allocated different bits to transmit in the binary expansion of the number to be communicated. The optimal scheme suggests allocating $\frac{1}{2}$ of the senders to the first bit, $\frac{1}{4}$ to the second bit etc.

channel can compensate for transmission error so as to still be more accurate than the error free Boolean channel. Indeed to compensate for an error rate less than 4.5% requires only one additional sender. Regarding robustness to ignorance concerning the input distribution our results are rather more mixed. If reality is well modelled by the family of symmetric beta distributions then across all possible parameter values there is an upper bound on the minimal number of senders required by the vague channel. On the other hand, no such upper bound exists for the general family of beta distributions. This is mainly because an asymmetric model of this kind allows for the case that reality may turn out to be particularly favourable for the Boolean channel. For example, this is the case if the distribution on inputs is heavily peaked at either $\frac{1}{4}$ or $\frac{3}{4}$.

In the current paper we have focussed on vague labels with fixed definitions. However, a common feature of adjectives in natural language is that they are context dependent. For example, the description *short* has a different meaning when applied to the restricted class of basketball players than to the general class of potential suspects in the Bristol robbery. One potential mechanism by which relative descriptors of this kind could be incorporated into the current model would be for both speakers and listeners to employ a form of context dependent scaling. For instance, suppose that z is the underlying variable to be communicated, e.g. unscaled height in the robber example, and further suppose that for a reference class C, z has the distribution function F_C. If both senders and receivers have sufficient knowledge of z on class C to have a good estimate of F_C, then channels of the following form can be defined by employing rescaling. Senders evaluate the scaled variable $x = F_C(z)$, which is uniformly distributed on [0, 1] provided that inputs are restricted to the class C. This can then be transmitted using vague channels of the form proposed above, with the receiver obtaining an estimate y of x, which they then rescale according to $F_C^{-1}(y)$ in order to give an estimate of z. In this case the production function for the input z is $P(S = 1|z) = \mu_{L_2}(F_C(z))$, and Fig. 14 illustrates this scaling process for the reference classes 'UK males', perhaps the reference class for the robbery example, and 'Basketball players'. In particular, Fig. 14c shows the membership functions for *tall* in the two different contexts. Additional work is required to investigate the efficacy of this approach from a communication perspective.[13]

In addition to signalling errors as discussed in Sect. 6, there are two additional sources of noise that will naturally occur for the type of communication channel we have proposed. Firstly, assuming a distributed learning model in which individuals infer the meanings of labels from repeated experiences of language use, it is inevitable that there will be variations in definitions between individuals. Secondly, we have assumed that all senders are describing the same input value. In reality this sensory data is likely to be subject to noise from a variety of sources. A future challenge is then to undertake a comparative study of vague and Boolean channels in the presence of both types of noise.

[13] In the "Appendix", Example 9 illustrates how this model of context can be applied so as to give an account of absolute adjectives.

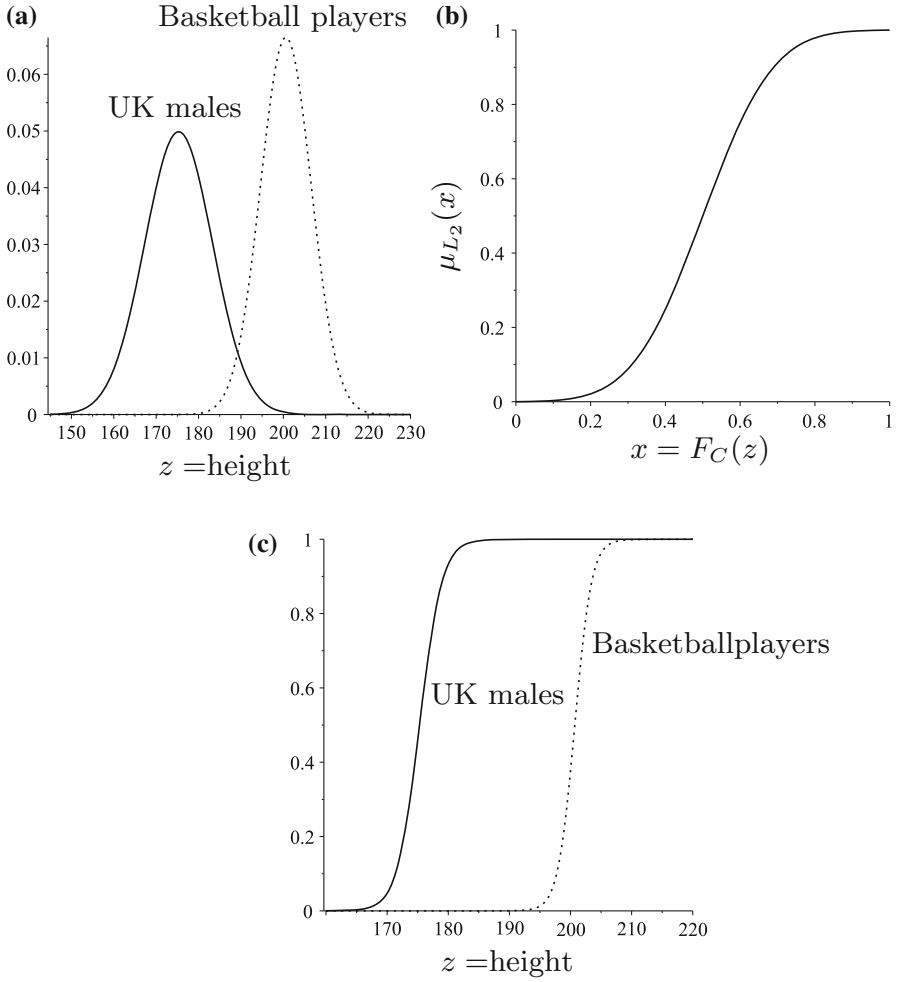

Fig. 14 Context scaling for the reference classes 'UK males' and 'Basketball players'. **a** Distributions of two reference classes. **b** Optimal membership for L_2 given 8 senders. **c** Scaled production functions $P(S = 1|z) = \mu_{L_2}(F_C(z))$ for the two reference classes. These correspond to the membership functions for tall in the two different contexts

In summary, the results presented in this paper suggest that vagueness acting as a source of randomness in assertion decisions, can be useful in communication scenarios where the number of relevant description words is moderately large and when there is aggregation of signals from several senders. However, the extent to which such scenarios occur in natural language and whether or not they are sufficiently common to explain the ubiquitousness of vague terms, remains very much an open question.

Acknowledgements The authors would like to thank the anonymous referees for their insightful comments and suggestions. All underlying data is included in full within the paper.

Appendix

Proof of Theorem 1 This is a special case of Theorem 2. □

Proof of Theorem 2 For fixed $\theta_i : i = 0, \ldots, k$ then we need to pick y_i so as to minimize

$$\mathbb{E}^{\mathbf{B}}((x-y)^2) = \sum_{i=1}^{k} \int_{\theta_{i-1}}^{\theta_i} (x - y_i)^2 \, \mathrm{d}x$$

This corresponds to selecting y_i so as to minimize $\int_{\theta_{i-1}}^{\theta_i} (x - y_i)^2 \, \mathrm{d}x$ for $i = 1, \ldots, k$. Now,

$$\int_{\theta_{i-1}}^{\theta_i} (x - y_i)^2 \, \mathrm{d}x = \frac{1}{3}\theta_i^3 - \theta_i^2 y_i + \theta_i y_i^2 - \frac{1}{3}\theta_{i-1}^3 + \theta_{i-1}^2 y_i - \theta_{i-1} y_i^2$$

Now differentiating with respect to y_i is gives us the following:

$$\frac{\partial \mathbb{E}^{\mathbf{B}}((x-y)^2)}{\partial y_i} = -\theta_i^2 + 2y_i\theta_i + \theta_{i-1}^2 - 2\theta_{i-1}y_i$$

Hence, setting $\frac{\partial \mathbb{E}^{\mathbf{B}}((x-y)^2)}{\partial y_i} = 0$ gives us that

$$2y_i(\theta_i - \theta_{i-1}) = \theta_i^2 - \theta_{i-1}^2 \Rightarrow y_i = \frac{\theta_i + \theta_{i-1}}{2}$$

Also, since the second derivative $\frac{\partial^2 \mathbb{E}^{\mathbf{B}}((x-y)^2)}{\partial y_i^2} = 2(\theta_i - \theta_{i-1}) > 0$ then this corresponds to the minimum. Hence, substituting we obtain the following expression:

$$\mathbb{E}^{\mathbf{B}}((x-y)^2) = \sum_{i=1}^{n} \frac{1}{12}\theta_i^3 - \frac{1}{4}\theta_i^2\theta_{i-1} + \frac{1}{4}\theta_i\theta_{i-1}^2 - \frac{1}{12}\theta_{i-1}^3$$

Differentiating with respect to θ_i we obtain the following:

$$\frac{\partial \mathbb{E}^{\mathbf{B}}((x-y)^2)}{\partial \theta_i} = -\frac{1}{2}\theta_i\theta_{i-1} + \frac{1}{4}\theta_{i-1}^2 - \frac{1}{4}\theta_{i+1}^2 + \frac{1}{2}\theta_{i+1}\theta_i$$

Hence, setting $\frac{\partial \mathbb{E}^{\mathbf{B}}((x-y)^2)}{\partial t_i} = 0$ gives us that

$$\frac{1}{2}\theta_{i+1}\theta_i - \frac{1}{2}\theta_i\theta_{i-1} = \frac{1}{4}\theta_{i+1}^2 - \frac{1}{4}\theta_{i-1}^2 \Rightarrow \frac{\theta_i(\theta_{i+1}-\theta_{i-1})}{2} = \frac{(\theta_{i+1}+\theta_{i-1})(\theta_{i+1}-\theta_{i-1})}{4}$$
$$\Rightarrow \theta_{i+1} = 2\theta_i - \theta_{i-1}.$$

Now since $\theta_0 = 0$ and $\theta_k = 1$ then this recurrence relation has the closed form $\theta_i = \frac{i}{k}$. Finally now that since the second derivative $\frac{\partial^2 \mathbb{E}^{\mathbf{B}}((x-y)^2)}{\partial \theta_i^2} = \frac{1}{2}(\theta_{i+1} - \theta_{i-1}) > 0$ then this is a minimum. □

Proof of Lemma 3 If $x \in [\frac{i-1}{k-1}, \frac{i}{k-1})$ then $P(L_r|x) = 0$ for $r > i+1$ and $r < i$. Hence, $n = n_i + n_{i-1}$ and $\sum_{j=1}^n R_j = i = (i)n_i + (i-1)n_{i-1}$. Therefore,

$$\frac{\sum_{j=1}^n R_j}{n(k-1)} = \frac{(i)n_i + (i-1)n_{i-1}}{n(k-1)} = \frac{(i)n_i + (i-1)(n-n_i)}{n(k-1)}$$
$$= \frac{n_i + n(i-1)}{n(k-1)} = \frac{\frac{n_i}{n} + i - 1}{k-1} = \frac{\frac{n_i}{n_{i-1}+n_i} + i - 1}{k-1}$$

Now for $x \in [\frac{i-1}{k-1}, \frac{i}{k-1}), n_i$ has a binomial distribution with probability parameter $P(S_j = i|x) = \mu_{L_{i+1}}(x) = (k-1)x - (i-1)$. Hence,

$$\mathbb{E}^{\mathbf{V}}(y|x) = \mathbb{E}\left(\frac{\frac{n_i}{n} + i - 1}{k-1}|x\right) = \frac{\frac{1}{n}\mathbb{E}(n_i|x) + i - 1}{k-1} = \frac{(k-1)x - (i-1) + i - 1}{k-1} = x$$

□

Proof of Theorem 4 Now,

$$\mathbb{E}^{\mathbf{B}}((x-y)^2) = \int_0^1 \mathbb{E}^{\mathbf{B}}((x-y)^2|x)\,\mathrm{d}x = \sum_{i=1}^k \int_{\frac{i-1}{k}}^{\frac{i}{k}} \mathbb{E}^{\mathbf{B}}((x-y)^2|x)\,\mathrm{d}x$$
$$= \sum_{i=1}^k \int_{\frac{i-1}{k}}^{\frac{i}{k}} \left(x - \frac{2i-1}{2k}\right)^2 \mathrm{d}x = k\left(\frac{1}{12k^3}\right) = \frac{1}{12k^2}$$

Also,

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) = \int_0^1 \mathbb{E}^{\mathbf{V}}((x-y)^2|x)\,\mathrm{d}x = \sum_{i=1}^k \int_{\frac{i-1}{k-1}}^{\frac{i}{k-1}} \mathbb{E}^{\mathbf{V}}((x-y)^2|x)\,\mathrm{d}x$$

Now by Lemma 3 $\mathbb{E}^{\mathbf{V}}(y|x) = x$ and hence

$$\mathbb{E}^{\mathbf{V}}((x-y)^2|x) = \mathbb{E}^{\mathbf{V}}((\mathbb{E}^{\mathbf{V}}(y|x) - y)^2|x) = \mathbb{V}^{\mathbf{V}}(y|x)$$

Also, for $x \in [\frac{i-1}{k-1}, \frac{i}{k-1}), n_i$ is binomially distributed with parameters n and $\mu_{L_{i+1}}(x) = x(k-1) - (i-1)$. Hence,

$$\begin{aligned}\mathbb{E}^{\mathbf{V}}((x-y)^2|x) &= \mathbb{V}\left(\frac{\frac{n_i}{n}+i-1}{k-1}|x\right) = \mathbb{V}\left(\frac{n_i}{n(k-1)}|x\right)\\ &= \frac{1}{n^2(k-1)^2}\mathbb{V}(n_i|x) = \frac{(x(k-1)-(i-1))(i-x(k-1))}{(k-1)^2 n}\end{aligned}$$

so that

$$\int_{\frac{i-1}{k-1}}^{\frac{i}{k-1}} \mathbb{E}^{\mathbf{V}}((x-y)^2|x)\,\mathrm{d}x = \frac{1}{6n(k-1)^3}$$

and therefore,

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) = (k-1)\frac{1}{6n(k-1)^3} = \frac{1}{6n(k-1)^2}$$

In contrast, we have for the k symbol Boolean channel that:

$$\mathbb{E}^{\mathbf{B}}((x-y)^2) = k\int_{\frac{i-1}{k}}^{\frac{i}{k}} \left(x-\frac{2i-1}{2k}\right)^2 \mathrm{d}x = k\left(\frac{1}{12k^3}\right) = \frac{1}{12k^2}$$

Consequently, $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ if and only if $\frac{1}{6n(k-1)^2} \leq \frac{1}{12k^2}$ if and only if $n \geq \frac{2k^2}{(k-1)^2}$ as required. □

Proof of Theorem 5 For $x \in [0,1]$ we have that:

$$\begin{aligned}P(R_j = 1|x) &= P(R_j = 1|S_j = 1)P(S_j = 1|x) + P(R_j = 1|S_j = 0)P(S_j = 0|x)\\ &= (1-\alpha)x + \alpha(1-x)\end{aligned}$$

Hence, $T = \sum_{j=1}^{n} R_j$ is binomially distributed with mean $n((1-\alpha)x + \alpha(1-x))$ and variance $n((1-\alpha)x + \alpha(1-x))(\alpha x + (1-\alpha)(1-x))$. Now

$$\mathbb{E}^{\mathbf{V}}((x-y)^2|x,\alpha) = \mathbb{E}^{\mathbf{V}}(y^2|x,\alpha) - 2x\mathbb{E}^{\mathbf{V}}(y|x,\alpha) + x^2$$

Here we have that:

$$\mathbb{E}^{\mathbf{V}}(y|x,\alpha) = \frac{1}{n}\mathbb{E}(T|x,\alpha) = (1-\alpha)x + \alpha(1-x)$$

and

$$\begin{aligned}\mathbb{E}^{\mathbf{V}}(y^2|x,\alpha) &= \frac{1}{n^2}\mathbb{E}(T^2|x,\alpha) = \frac{1}{n^2}(\mathbb{V}(T|x,\alpha) + \mathbb{E}(T|x,\alpha)^2)\\ &= \frac{((1-\alpha)x + \alpha(1-x))(\alpha x + (1-\alpha)(1-x))}{n} + ((1-\alpha)x + \alpha(1-x))^2\end{aligned}$$

Substituting and expanding we obtain:

$$\mathbb{E}^{\mathbf{V}}(y^2|x,\alpha) = \frac{1}{n}\left(4\alpha x^2 + x - x^2 - 4\alpha x - 4\alpha^2 x^2 + 4\alpha^2 x + \alpha - \alpha^2 + 4n\alpha^2 x^2 - 4n\alpha^2 x + n\alpha^2\right)$$

Hence,

$$\mathbb{E}^{\mathbf{V}}((x-y)^2|\alpha) = \int_0^1 \mathbb{E}^{\mathbf{V}}((x-y)^2|x,\alpha)\,\mathrm{d}x = \frac{1}{6n}\left(2\alpha - 2\alpha^2 + 2n\alpha^2 + 1\right)$$

Hence, trivially $\mathbb{E}^{\mathbf{V}}((x-y)^2|\alpha)$ is a decreasing function of n and $\lim_{n\to\infty} \mathbb{E}^{\mathbf{V}}((x-y)^2|\alpha) = \frac{1}{3}\alpha^2$. Now for $\alpha \geq \frac{1}{4}, \frac{1}{3}\alpha^2 \geq \frac{1}{48} = \mathbb{E}^{\mathbf{B}}((x-y)^2)$ and hence $\forall n \geq 1\ \mathbb{E}^{\mathbf{V}}((x-y)^2|\alpha) > \mathbb{E}^{\mathbf{B}}((x-y)^2)$. In the case that $\alpha < \frac{1}{4}$ we require that:

$$\frac{1}{6n}\left(2\alpha - 2\alpha^2 + 2n\alpha^2 + 1\right) \leq \frac{1}{48} \Rightarrow n \geq \frac{8(2\alpha - 2\alpha^2 + 1)}{1 - 16\alpha^2}$$

as required. □

Proof of Theorem 6 This is a special case of Theorem 7. □

Proof of Theorem 7 Recall from Sect. 5 we have for the vague channel that

$$\mathbb{E}^{\mathbf{V}}((x-y)^2|x) = \frac{x(1-x)}{n}$$

Hence,

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) = \int_0^1 \frac{x - x^2 x^{s-1}(1-x)^{t-1}}{n\qquad \beta(s,t)}\mathrm{d}x = \frac{1}{n\beta(s,t)}(\beta(s+1,t) - \beta(s+2,t))$$

Now,

$$\frac{\beta(s+1,t)}{\beta(s,t)} = \frac{s}{s+t} \quad \text{and} \quad \frac{\beta(s+2,t)}{\beta(s,t)} = \frac{(s+1)t}{(s+t+1)(s+t)}$$

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) = \frac{1}{n}\left(\frac{s}{s+t} - \frac{(s+1)s}{(s+t+1)(s+t)}\right) = \frac{st}{n(s+t+1)(s+t)}$$

For the Boolean channel we have that

$$\mathbb{E}^{\mathbf{B}}((x-y)^2) = \int_0^{0.5} \left(x - \frac{1}{4}\right)^2 \frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x + \int_{0.5}^1 \left(x - \frac{3}{4}\right)^2 \frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x$$

Now consider:

$$\int_0^{0.5}\left(x-\frac{1}{4}\right)^2\frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x=\int_0^{0.5}\left(x^2-\frac{1}{2}x+\frac{1}{16}\right)\frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x$$
$$=\int_0^{0.5}\frac{x^{s+1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x-\frac{1}{2}\int_0^{0.5}\frac{x^{s}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x$$
$$+\frac{1}{16}\int_0^{0.5}\frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x$$
$$=\frac{\beta\left(\frac{1}{2};s+2,t\right)}{\beta(s,t)}-\frac{1}{2}\frac{\beta\left(\frac{1}{2};s+1,t\right)}{\beta(s,t)}+\frac{1}{16}\frac{\beta\left(\frac{1}{2};s,t\right)}{\beta(s,t)}$$

Now consider $\frac{\beta\left(\frac{1}{2};s+2,t\right)}{\beta(s,t)}$ and recall that

$$\beta(s+2,t)=\frac{(s+1)s}{(s+t+1)(s+t)}\beta(s,t)\Rightarrow\frac{\beta\left(\frac{1}{2};s+2,t\right)}{\beta(s,t)}$$
$$=\frac{s(s+1)}{(s+t)(s+t+1)}\frac{\beta\left(\frac{1}{2};s+2,t\right)}{\beta(s+2,t)}=\frac{s(s+1)}{(s+t)(s+t+1)}I_{\frac{1}{2}}(s+2,t)$$

Furthermore,

$$I_{\frac{1}{2}}(s+2,t)=I_{\frac{1}{2}}(s+1,t)-\frac{\left(\frac{1}{2}\right)^{s+t+1}}{(s+1)\beta(s+1,t)}$$
$$=I_{\frac{1}{2}}(s+1,t)-\frac{\left(\frac{1}{2}\right)^{s+t+1}(s+t)}{s(s+1)\beta(s,t)}\quad\text{since}\quad\beta(s+1,t)=\frac{s}{s+t}\beta(s,t)$$

Also, we have that:

$$I_{\frac{1}{2}}(s+1,t)=I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}}{s\beta(s,t)}$$

Hence,

$$I_{\frac{1}{2}}(s+2,t)=I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}}{s\beta(s,t)}-\frac{\left(\frac{1}{2}\right)^{s+t+1}(s+t)}{s(s+1)\beta(s,t)}$$
$$=I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}}{s\beta(s,t)}\left(\frac{\frac{3}{2}s+\frac{1}{2}t+1}{s+1}\right)$$

Therefore, by substituting we obtain

$$\frac{\beta\left(\frac{1}{2};s+2,t\right)}{\beta(s,t)}=\frac{s(s+1)}{(s+t)(s+t+1)}\left(I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}}{s\beta(s,t)}\left(\frac{\frac{3}{2}s+\frac{1}{2}t+1}{s+1}\right)\right)$$
$$=\frac{s(s+1)}{(s+t)(s+t+1)}I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}\left(\frac{3}{2}s+\frac{1}{2}t+1\right)}{(s+t)(s+t+1)\beta(s,t)}$$

Now consider

$$\frac{\beta\left(\frac{1}{2};s+1,t\right)}{\beta(s,t)}=\frac{s}{s+t}I_{\frac{1}{2}}(s+1,t)\quad\text{since}\quad\beta(s+1,t)=\frac{s}{s+t}\beta(s,t)$$

Furthermore,

$$I_{\frac{1}{2}}(s+1,t)=I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}}{s\beta(s,t)}$$

Hence,

$$\frac{\beta\left(\frac{1}{2};s+1,t\right)}{\beta(s,t)}=\frac{s}{s+t}\left(I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}}{s\beta(s,t)}\right)=\frac{s}{s+t}I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}}{(s+t)\beta(s,t)}$$

Finally,

$$\frac{\beta\left(\frac{1}{2};s,t\right)}{\beta(s,t)}=I_{\frac{1}{2}}(s,t)$$

Therefore;

$$\begin{aligned}\int_0^{0.5}\left(x-\frac{1}{4}\right)^2\frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x&=\frac{s(s+1)}{(s+t)(s+t+1)}I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}\left(\frac{3}{2}s+\frac{1}{2}t+1\right)}{(s+t)(s+t+1)\beta(s,t)}\\&-\frac{1}{2}\left(I_{\frac{1}{2}}(s,t)-\frac{\left(\frac{1}{2}\right)^{s+t}}{s\beta(s,t)}\right)+\frac{1}{16}I_{\frac{1}{2}}(s,t)\\&=\frac{1}{(s+t)(s+t+1)}\left(\frac{I_{\frac{1}{2}}(s,t)(9s^2+9s-6st+t^2+t)}{16}-\frac{\left(\frac{1}{2}\right)^{s+t+1}(2s+1)}{\beta(s,t)}\right)\end{aligned}$$

Now consider

$$\int_{0.5}^{1}\left(x-\frac{3}{4}\right)^2\frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x=\int_0^{0.5}\left(y-\frac{1}{4}\right)^2\frac{y^{t-1}(1-y)^{s-1}}{\beta(t,s)}\mathrm{d}y$$

by substituting $y=1-x$ and since $\beta(s,t)=\beta(t,s)$. Hence, by exchanging s and t in the previous expression we have that:

$$\begin{aligned}&\int_{0.5}^{1}\left(x-\frac{3}{4}\right)^2\frac{x^{s-1}(1-x)^{t-1}}{\beta(s,t)}\mathrm{d}x\\&=\frac{1}{(s+t)(s+t+1)}\left(\frac{I_{\frac{1}{2}}(t,s)(9t^2+9t-6st+s^2+s)}{16}-\frac{\left(\frac{1}{2}\right)^{s+t+1}(2t+1)}{\beta(s,t)}\right)\\&=\frac{1}{(s+t)(s+t+1)}\left(\frac{\left(1-I_{\frac{1}{2}}(s,t)\right)(9t^2+9t-6st+s^2+s)}{16}-\frac{\left(\frac{1}{2}\right)^{s+t+1}(2t+1)}{\beta(s,t)}\right)\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}^{\mathbf{B}}((x-y)^2) = &\frac{1}{(s+t)(s+t+1)}\left(\frac{I_{\frac{1}{2}}(s,t)(9s^2+9s-6st+t^2+t)}{16} - \frac{\left(\frac{1}{2}\right)^{s+t+1}(2s+1)}{\beta(s,t)}\right)\\ &+\frac{1}{(s+t)(s+t+1)}\left(\frac{(1-I_{\frac{1}{2}}(s,t))(9t^2+9t-6st+s^2+s)}{16} - \frac{\left(\frac{1}{2}\right)^{s+t+1}(2t+1)}{\beta(s,t)}\right)\\ &= \frac{1}{(s+t)(s+t+1)}\left(\frac{1}{2}I_{\frac{1}{2}}(s,t)(s-t)(s+t+1)\right.\\ &\left.-\frac{\left(\frac{1}{2}\right)^{s+t}(s+t+1)}{\beta(s,t)} + \frac{9t^2+9t-6st+s^2+s}{16}\right)\end{aligned}$$

Therefore, for $\mathbb{E}^{\mathbf{V}}((x-y)^2) \leq \mathbb{E}^{\mathbf{B}}((x-y)^2)$ we require that

$$\begin{aligned}\frac{st}{n(s+t+1)(s+t)} \leq &\frac{1}{(s+t)(s+t+1)}\left(\frac{1}{2}I_{\frac{1}{2}}(s,t)(s-t)(s+t+1)\right.\\ &\left.-\frac{\left(\frac{1}{2}\right)^{s+t}(s+t+1)}{\beta(s,t)} + \frac{9t^2+9t-6st+s^2+s}{16}\right)\end{aligned}$$

Hence,

$$n \geq \frac{16\beta(s,t)st}{8\beta\left(\frac{1}{2};s,t\right)(s-t)(s+t+1) - 16\left(\frac{1}{2}\right)^{s+t}(s+t+1) + (9t^2+9t-6st+s^2+s)\beta(s,t)}$$

as required. □

Proof of Theorem 8 A vague channel is defined such that $L_1 = [0, \theta)$ and $L_2 = [\theta, 1]$ where θ is a random variable in [0, 1] with density function f and cumulative distribution function F, so that:

$$P(S=0) = P(x<\theta) = 1 - F(x) \quad \text{and} \quad P(S=1) = P(x \geq \theta) = F(x)$$

We assume that the density f is continuous on [0, 1] and hence $\mathbb{E}(\theta) \in (0, 1)$ and $\mathbb{V}(\theta) > 0$. For any such channel the error minimising estimator of x from T is:

$$y = \mathbb{E}(x|T) = \frac{\int_0^1 xP(T|x)\,\mathrm{d}x}{\int_0^1 P(T|x)\,\mathrm{d}x}$$

When there is only one sender then either $T = 1$ or $T = 0$. Now $P(T=1|x) = P(S=1|x) = F(x)$ and hence:

$$y_1 = \mathbb{E}(x|T=1) = \frac{\int_0^1 xF(x)\,\mathrm{d}x}{\int_0^1 F(x)\,\mathrm{d}x}$$

Now

$$\int_0^1 F(x)\,\mathrm{d}x = \int_0^1 \int_0^x f(\theta)\,\mathrm{d}\theta\,\mathrm{d}x = \int_0^1 f(\theta) \int_\theta^1 \mathrm{d}x\,\mathrm{d}\theta$$
$$= \int_0^1 f(\theta)(1-\theta)\,\mathrm{d}\theta = \mathbb{E}(1-\theta) = 1 - \mathbb{E}(\theta)$$

Also

$$\int_0^1 xF(x)\,\mathrm{d}x = \int_0^1 x \int_0^x f(\theta)\,\mathrm{d}\theta\,\mathrm{d}x = \int_0^1 f(\theta) \int_\theta^1 x\,\mathrm{d}x\,\mathrm{d}\theta$$
$$= \int_0^1 f(\theta)\frac{(1-\theta^2)}{2}\,\mathrm{d}\theta = \frac{1}{2}\mathbb{E}(\theta^2) = \frac{1}{2}(1 - \mathbb{E}(\theta^2))$$

Hence,

$$y_1 = \frac{\frac{1}{2}(1 - \mathbb{E}(\theta^2))}{1 - \mathbb{E}(\theta)}$$

Similarly, $P(T = 0|x) = P(S = 0|x) = 1 - F(x)$ and hence:

$$y_0 = \mathbb{E}(x|T) = \frac{\int_0^1 x(1 - F(x))\,\mathrm{d}x}{\int_0^1 1 - F(x)\,\mathrm{d}x}$$

Now,

$$\int_0^1 1 - F(x)\,\mathrm{d}x = 1 - \int_0^1 F(x)\,\mathrm{d}x = \mathbb{E}(\theta)$$

Also,

$$\int_0^1 x(1 - F(x))\,\mathrm{d}x = \int_0^1 x\,\mathrm{d}x - \int_0^1 xF(x)\,\mathrm{d}x = \frac{1}{2} - \frac{1}{2}(1 - \mathbb{E}(\theta^2)) = \frac{1}{2}\mathbb{E}(\theta^2)$$

Hence,

$$y_0 = \frac{\frac{1}{2}\mathbb{E}(\theta^2)}{\mathbb{E}(\theta)}$$

From this we have that:

$$\mathbb{E}^{\mathbf{V}}((x-y)^2|x) = (x-y_1)^2F(x) + (x-y_0)^2(1-F(x))$$

and hence,

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) = \int_0^1 (x-y_1)^2F(x)\,\mathrm{d}x + \int_0^1 (x-y_0)^2(1-F(x))\,\mathrm{d}x$$

Now,

$$\begin{aligned}\int_0^1 (x-y_1)^2F(x)\,\mathrm{d}x &= \int_0^1 (x^2-2y_1x+y_1^2)F(x)\,\mathrm{d}x\\ &= \int_0^1 x^2F(x)\,\mathrm{d}x - 2y_1\int_0^1 xF(x)\,\mathrm{d}x + y_1^2\int_0^1 F(x)\,\mathrm{d}x\end{aligned}$$

Also,

$$\begin{aligned}\int_0^1 (x-y_0)^2(1-F(x))\,\mathrm{d}x &= \int_0^1 (x^2-2y_0x+y_0^2)(1-F(x))\,\mathrm{d}x\\ &= \int_0^1 x^2-2y_0x+y_0^2\,\mathrm{d}x - \int_0^1 (x^2-2y_0x+y_0^2)F(x)\,\mathrm{d}x\\ &= \frac{1}{3} - y_0 + y_0^2 - \int_0^1 x^2F(x)\,\mathrm{d}x + 2y_0\int_0^1 xF(x)\,\mathrm{d}x - y_0^2\int_0^1 F(x)\,\mathrm{d}x\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}^{\mathbf{V}}((x-y)^2) &= \int_0^1 x^2F(x)\,\mathrm{d}x - 2y_1\int_0^1 xF(x)\,\mathrm{d}x + y_1^2\int_0^1 F(x)\,\mathrm{d}x\\ &\quad + \frac{1}{3} - y_0 + y_0^2 - \int_0^1 x^2F(x)\,\mathrm{d}x + 2y_0\int_0^1 xF(x)\,\mathrm{d}x - y_0^2\int_0^1 F(x)\,\mathrm{d}x\\ &= \frac{1}{3} - y_0 + y_0^2 + (y_0-y_1)2\int_0^1 xF(x)\,\mathrm{d}x + (y_1^2-y_0^2)\int_0^1 xF(x)\,\mathrm{d}x\\ &= \frac{1}{3} - y_0 + y_0^2 + (y_0-y_1)(1-\mathbb{E}(\theta^2)) + (y_1^2-y_0^2)(1-\mathbb{E}(\theta))\end{aligned}$$

Letting $w = \mathbb{E}(\theta^2)$ and $\mathbb{E}(\theta) = z$ we have that $y_0 = \frac{1}{2}\left(\frac{w}{z}\right), y_0 = \frac{1}{2}\left(\frac{1-w}{1-z}\right)$ and by substituting that:

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) = \frac{1}{3} - \frac{1}{4}\frac{w^2+z-2wz}{z(1-z)}$$

Now since f is continuous on [0, 1] it follows that $0<z^2<w<z<1$. From this we have that $\frac{w^2+z-2wz}{z(1-z)} > 0$. Hence, $\mathbb{E}^{\mathbf{V}}((x-y)^2)$ is minimal when $\frac{w^2+z-2wz}{z(1-z)}$ is maximal. Now,

$$\frac{\partial}{\partial w}\frac{w^2+z-2wz}{z(1-z)} = \frac{2(w-z)}{z(1-z)} < 0$$

Hence, since $w > z^2$ it therefore follows that:

$$\frac{w^2+z-2wz}{z(1-z)} < \left.\frac{w^2+z-2wz}{z(1-z)}\right|_{w=z^2} = -z^2+z+1$$

Furthermore $-z^2+z+1$ has a maximum at $z=\frac{1}{2}$ and hence

$$\frac{w^2+z-2wz}{z(1-z)} < -z^2+z+1 \leq \left.-z^2+z+1\right|_{z=\frac{1}{2}} = \frac{5}{4}$$

Hence,

$$\mathbb{E}^{\mathbf{V}}((x-y)^2) > \frac{1}{3} - \frac{1}{4}\left(\frac{5}{4}\right) = \frac{1}{48} = \mathbb{E}^{\mathbf{B}}((x-y)^2)$$

□

Example 9 Kennedy (2006) identifies the class of *absolute* gradable adjectives which, while relating to an underlying bounded scale, refer specifically to the end points of that scale and are either crisp or a least much less vague than terms such as short or tall. In this example we illustrate how the model of context outlined in Sect. 9 can be used to give an account of absolute adjectives. For example, consider the question 'is the glass full?'. Here the underlying variable z described by the adjective *full* is something like the ratio of the volume of liquid in the glass over the total volume of the glass, and the context relates to the glasses of liquid typically encountered. Now suppose that the distribution of z in this context is skewed towards the upper bound 1. For instance, Fig. 15a shows the density function for this context assuming that z is distributed according to a beta distribution with parameter values $s=7$ and $t=1$. In this case, for a channel optimised for 2 senders (Fig. 15b), the production function for z, i.e. the membership function of *full*, will be close to a step function near 1 (Fig. 15c). In summary, this example shows that for a pair of labels, in a context in which the underlying scale is bounded and the distribution is skewed towards the upper (lower) bound on that scale, if the production function of x is optimised for a small number of senders then the resulting membership function for the second (first) label is 'almost' crisp and with the transition between 0 and 1 membership occurring close to the upper (lower) boundary on the scale. This account seems to have similarities with the probabilistic treatment of

absolute adjectives in Qing and Franke (2014), although in the latter the optimisation of production functions has a different motivation.

One criticism of this approach is the assumption of an asymmetric distribution on z which, while providing some justification for an almost crisp membership for *full*, does not naturally generate a similar membership function for *empty* as being close to a step function near 0. An alternative would be to assume a symmetric distribution for z which is peaked at both of the scale boundaries. In other words, to assume that there is a bias towards glasses that are either very full or very empty. One possibility is to consider symmetric beta distributions with $s = t < 1$. For example, Fig. 16a shows the density function for the beta distribution with $s = t = 0.2$. However, such distributions are not in themselves sufficient to generate intuitive membership functions for *empty* and *full*. The problem lies in our having limited ourselves to only two labels and in the fact that senders must choose exactly

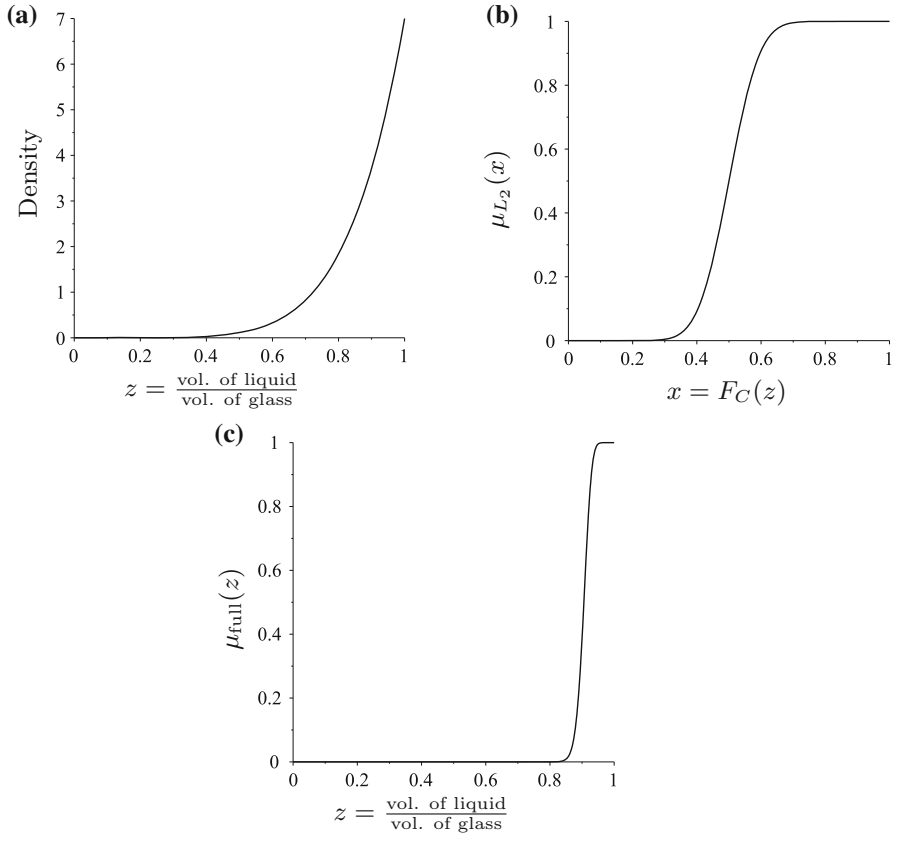

Fig. 15 Context scaling for the reference class 'glass of water' for underling variable $z = \frac{\text{vol. of liquid}}{\text{vol. of glass}}$ where z is distributed according to a beta distribution with parameters $s = 7$ and $t = 1$. **a** Distribution of the reference class. **b** Optimal membership for L_2 given 2 senders. **c** Scaled production function corresponding to the membership function of full

one of these to transmit given any input value. Consequently, for all values of z, including those between 0 and 1, it holds that $\mu_{\text{empty}}(z) + \mu_{\text{full}}(z) = 1$. Given this constraint it follows, for example, that assuming a symmetric beta distribution on z with $s = t$, in the limit as s tends to 0 we obtain membership functions for *empty* and *full* such that:

$$\mu_{\text{empty}}(z) = \begin{cases} 1 : z = 0 \\ 0.5 : 0 < z < 1 \\ 0 : z = 1 \end{cases} \quad \text{and} \quad \mu_{\text{full}}(z) = \begin{cases} 0 : z = 0 \\ 0.5 : 0 < z < 1 \\ 1 : z = 1 \end{cases}$$

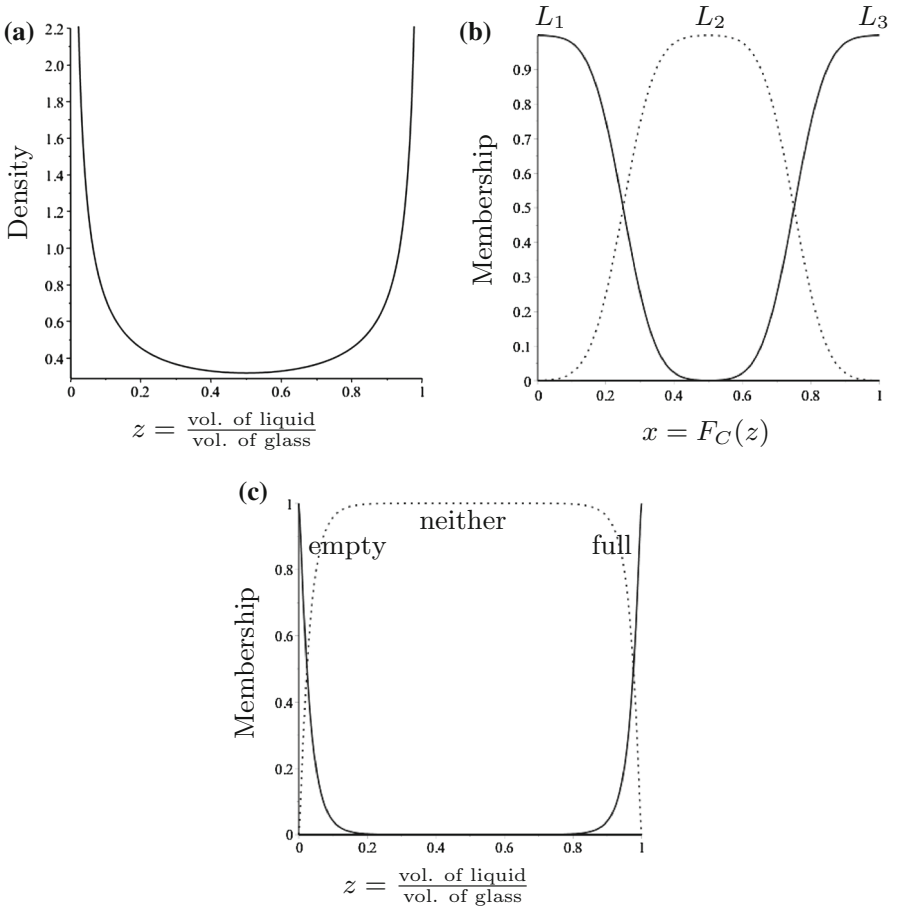

Fig. 16 Context scaling for the reference class 'glass of water' for underling variable $z = \frac{\text{vol. of liquid}}{\text{vol. of glass}}$ where z is distributed according to a symmetric beta distribution with parameters $s = 0.2$ and $t = 0.2$ and where the message set has three description labels. **a** Symmetric distribution of the reference class. **b** Gaussian membership functions for three labels. **c** Scaled production functions corresponding to the membership functions of empty, neither empty nor full and full

This is clearly counter intuitive since one would not expect absolute adjectives to have membership 0.5 for such a large range of values. A possible way around this is to add another label to the language, perhaps standing for 'neither empty nor full'. Figure 16b shows possible Gaussian membership functions for a general three label message set. Assuming that z is distributed according to a beta distribution with $s = t = 0.2$ then we obtain the membership functions for *empty*, *full* and *neither*, as shown in Fig. 16c.

References

Black, M. (1937). Vagueness: An exercise in logical analysis. *Philosophy of Science*, *4*(4), 427–455.

Cresswell, M. J. (1976). The semantics of degree. In B. Partee (Ed.), *Montague grammar* (pp. 261–292). London: Academic Press.

De Jaegher, K. (2003). A game-theoretic rationale for vagueness. *Linguistic and Philosophy*, *26*(5), 637–659.

D'Odorico, T., & Bennett, B. (2013). Automated reasoning on vague concepts using formal ontologies, with an application to event detection on video data. In *Commonsense 2013, Proceedings of 11th international symposium on logical formalizations of commonsense reasoning*, Aya Napa, Cyprus.

Dubois, D., & Prade, H. (1997). The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, *90*, 141–150.

Edgington, D. (1997). Vagueness by degrees. In R. Keefe & P. Smith (Eds.), *Vagueness: A reader* (pp. 294–316). Cambridge: MIT Press.

Egré, P., & Barberousse, A. (2014). Borel on the heap. *Erkenntnis*, *79*, 1043–1079.

Egré, P. (2016). Vague judgement: A probabilistic account. *Synthese*. doi:10.1007/s11229-016-1092-2.

Eyre, H., & Lawry, J. (2014). Language games with vague categories and negation. *Adaptive Behaviour*, *22*(5), 289–303.

Fine, K. (1975). Vagueness, truth and logic. *Synthese*, *30*, 265–300.

Franke, M., Jäger, G., & van Rooij, R. (2011). Vagueness, signalling and bounded rationality. In T. Onoda, D. Bekki & E. McCready (Eds.), *New frontiers in artificial intelligence. Lecture notes in computer science* (Vol. 6797, pp. 45–59). Berlin: Springer.

Grice, H. P. (1975). Logic in conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts* (pp. 41–58). London: Academic Press.

Gubner, J. A. (1993). Distributed estimation and quantization. *IEEE Transactions on Information Theory*, *39*(4), 1456–1459.

Hüllermeier, E. (2011). Fuzzy sets in machine learning and data mining. *Applied Soft Computing*, *11*, 1493–1505.

Hisdal, H. (1988). Are grades of membership probabilities. *Fuzzy Sets and Systems*, *25*, 325–348.

O'Connor, C. (2013). The evolution of vagueness. *Erkenntis*, *79*(4), 707–727.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge: MIT Press.

Kamp, H. (1975). Two theories of adjectives. In E. L. Keenan (Ed.), *Formal semantics of natural language* (pp. 123–155). Cambridge: Cambridge University Press.

Keefe, R., & Smith, P. (Eds.). (2002). *Vagueness: A reader*. Cambridge: MIT Press.

Kennedy, C. (2006). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45.

Kleene, S. C. (1952). *Introduction to metamathematics*. Princeton, New Jersey: D. Van Nostrand Company Inc.

Lassiter, D. (2011). Vagueness as probabilistic linguistic knowledge. In R. van Rooij & U. Sauerland (Eds.), *Vagueness in communication. Lecture notes in computer science* (Vol. 6517, pp. 127–150). Heidelberg: Springer.

Lassiter, D., & Goodman, N. D. (2013). Context scale structure, and statistics in the interpretation of positive-form adjectives. *Proceedings of SALT*, *23*, 587–610.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*. doi:10.1007/s11229-015-0786-1.

Lawry, J. (1998). A voting mechanism for fuzzy logic. *International Journal of Approximate Reasoning*, *19*(3), 315–333.

Lawry, J. (2008). Appropriateness measures: An uncertainty model for vague concepts. *Synthese*, *161*(2), 255–269.

Lawry, J., & Tang, Y. (2012). On truth-gaps, bipolar belief and the assertability of vague propositions. *Artificial Intelligence*, *191–192*, 20–41.

Lewis, D. (1969). *Convention: A philosophical study*. Cambridge: Harvard University Press.

Lipman, B. L. (2009). Why is language vague? Working paper, Department of Economics, Boston University.

Loginov, V. J. (1966). Probability treatment of Zadeh membership functions and their use in pattern recognition. *Engineering Cybernetics, 4*, 68–69.

Luo, Z.-Q. (2005). Universal decentralized estimation in a bandwidth constrained sensor network. *IEEE Transactions on Information Theory*, *51*(6), 2210–2219.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychology Review*, *63*, 81–97.

Qing, C., & Franke, M. (2014). Meaning and use of gradable adjectives: formal modeling meets empirical data. In *Proceedings of the 36'th annual conference of the cognitive science society (CogSci 2014)* (pp. 1204–1209).

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. *Proceedings of SALT*, *24*, 23–41.

Ribeiro, A., & Giannakis, G. B. (2006). Bandwidth-constrained distributed estimation for wireless sensor networks—Part I: Gaussian case. *IEEE Transactions on Signal Processing*, *54*(3), 1131–1143.

Smith, N. J. J. (2008). *Vagueness and degrees of truth*. Oxford: Oxford University Press.

Steels, L. (1997). The synthetic modelling of language origins. *Evolution of Communication*, *1*(1), 1–34.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study. *Behavioural and Brain Sciences*, *28*(4), 469–488.

van Deemter, K. (2009a). Utility and language generation: The case of vagueness. *Journal of Philosophical Logic*, *38*, 607–632.

van Deemter, K. (2009b). What game theory can do for NLG: The case of vague language. 12th European workshop on natural language generation (pp. 154–161).

Williamson, T. (1992). Vagueness and ignorance. *Proceedings of the Aristotlian Society*, *66*, 145–162.

Williamson, T. (1994). *Vagueness*. London: Routledge.

Xiao, J.-J., Cui, S., Luo, Z.-Q., & Goldsmith, A. J. (2006). Power scheduling of universal decentralized estimation in sensor networks. *IEEE Transactions on Signal Processing*, *54*(2), 413–422.